Explaining gift exchange – The limits of good intentions

Netzer, Nick ; Schmutzler, Armin

Abstract: This paper explores the limitations of intention-based social preferences as an explanation of gift-exchange between a firm and a worker. In a framework with one self-interested and one reciprocal player, gift-giving never arises in equilibrium. Instead, any equilibrium in a large class of multistage games must involve mutually unkind behavior of both players. Besides gift-exchange, this class of games also includes moral hazard models and the rotten kid framework. Even though equilibrium behavior may appear positively reciprocal in some of these games, the self-interested player never benefits from reciprocity. We discuss the relation of these results to the theoretical and empirical literature on gift-exchange in employment relations.

DOI: https://doi.org/10.1111/jeea.12086

Explaining Gift-Exchange – The Limits of Good Intentions*

Nick Netzer
University of Zürich

Armin Schmutzler
University of Zürich, CEPR and ENCORE

Third Revised Version
September 2013

Abstract

This paper explores the limitations of intention-based social preferences as an explanation of gift-exchange between a firm and a worker. In a framework with one self-interested and one reciprocal player, gift-giving never arises in equilibrium. Instead, any equilibrium in a large class of multi-stage games must involve mutually unkind behavior of both players. Besides gift-exchange, this class of games also includes moral hazard models and the rotten kid framework. Even though equilibrium behavior may appear positively reciprocal in some of these games, the self-interested player never benefits from reciprocity. We discuss the relation of these results to the theoretical and empirical literature on gift-exchange in employment relations.

Keywords: Reciprocity, Intentions, Moral Hazard, Gift-Exchange, Rotten Kid.

JEL Classification: C72, D03, D86, J01.

*Contact: Armin Schmutzler, University of Zürich, Department of Economics, Blümlisalpstr. 10, CH-8006 Zürich, Switzerland. Phone +41-44-634-2271, fax +41-44-634-4907, e-mail: armin.schmutzler@econ.uzh.ch. An earlier version of this paper was circulated under the title "Rotten Kids with Bad Intentions". We are grateful to the editor, Dirk Bergemann, five anonymous referees, as well as Daniel Benjamin, Paul Heidhues, Georg Kirchsteiger, Ulrike Malmendier, Matthew Rabin, Stefano DellaVigna, and seminar participants at the Universities of Frankfurt, Vienna and Zürich, UC Berkeley and UC Irvine, the ABEE symposium 2012, and the 2009 Meeting of German Economists Abroad for valuable suggestions. All errors are our own.

1 Introduction

Can firms profit from the social preferences and fairness concerns of their employees? An extensive literature, both theoretical and experimental, has answered in the affirmative. In an influential theoretical paper, Akerlof (1982) has proposed that a profit-maximizing firm may benefit from giving gifts to its employees, who acquire "sentiment for the firm" (p. 542) and reciprocate by providing more effort. A subsequent literature on worker-firm relations has applied various different theoretical models of social behavior that generate profitable gift-exchange.[1] Fehr, Kirchsteiger, and Riedl (1993) have first provided experimental evidence in favor of the gift-exchange hypothesis. They conclude that "...if workers behave reciprocally, paying high (fair) wages is not profit reducing but profit increasing" (p. 452), and they also find that reciprocity enhances efficiency. Similar results have later been found in many other experimental studies.[2]

This paper adds a more pessimistic view to the discussion. We analyze theoretically the interaction between one self-interested player, typically interpreted as a profit-maximizing firm, and one reciprocal player, typically interpreted as a worker. The worker is assumed to have intention-based social preferences (Rabin 1993, Dufwenberg and Kirchsteiger 2004). Models of intention-based social preferences incorporate the idea that agents benefit from reciprocating (or punishing) acts to which they attribute good (or bad) intentions.[3] Most contributions that deal with gift-exchange in employment relations mention intention-based social preferences as a possible justification, although few actually model them explicitly. Relying on the solution concept of an intentions-equilibrium, we show that any outcome in a large class of games must involve mutual unkindness, which implies that positive reciprocity is precluded. In a simple gift-exchange game, for instance, if the firm's gift were able to raise the worker's effort to a profitable level, then giving the gift would be in the firm's self-interest and no longer be associated with good intentions, contradicting the initial assumption that it raises the worker's effort. Ultimately, the equilibrium outcome must therefore involve no gift and no returned favor. Importantly, unkindness is an emerging equilibrium property and hence endogenous. Though players have ample opportunity to be kind, they choose not to.

[1] Rabin (1993) and Ruffle (1999) contain gift-exchange applications in psychological game frameworks (Geanakoplos, Pearce, and Stacchetti 1989, Battigalli and Dufwenberg 2009). Arbak and Kranich (2005) and Non (2012) apply the conceptual framework of Levine (1998) to the interaction between workers and firms. Englmaier and Leider (2012) and Bellemare and Shearer (2011) are additional models of reciprocity in employment relations. A general survey of reciprocity models is given by Sobel (2005).

[2] See Fehr, Goette, and Zehnder (2009) for a survey.

[3] Evidence suggests that social preferences in fact exhibit a strong intention-based component, see for instance Charness and Rabin (2002), Offerman (2002), Falk, Fehr, and Fischbacher (2003), McCabe, Rigdon, and Smith (2003), Falk and Fischbacher (2006), Falk, Fehr, and Fischbacher (2008) and Dohmen, Falk, Huffman, and Sunde (2009).

By the same argument, the equilibrium wage in a simple moral hazard game will never be kind. Interestingly, it will still be higher than theory predicts for a selfish worker. The high wage is, however, not paid to trigger the return of a favor, but to prevent the worker from punishing too strongly the firm's lack of good intentions. The payoff consequences are quite surprising. The player who benefits in material terms is the worker, while the firm loses. This stands in stark contrast to the view described in the beginning. The intuition is simple: Reciprocity increases the *elasticity* of efforts with respect to wages, because of the additional psychological payoffs the worker obtains from reacting to higher wages with higher effort. This makes larger wages attractive for the firm. However, reciprocity also has a negative *level* effect on efforts, because of the psychological motive to punish the unkind behavior that arises in equilibrium. This explains the negative effect on profits, and an ambiguous overall effect on efforts.

Our analysis also relates to the rotten kid theorem (Becker 1974, 1981), one of the earliest results in the theory of social preferences. Becker considers a framework where an egoistic player (the "rotten kid") can take an action that increases joint ("family") income, but reduces his own income, before an altruistic player (the "parent") makes a transfer to the kid. According to the rotten kid theorem, such one-sided altruism can induce an efficient outcome, because the kid can expect being rewarded for increasing family income. Benjamin (2010) has shown that the efficiency result continues to hold for a much broader class of outcome-based social preferences beyond altruism, including inequality aversion (Fehr and Schmidt 1999, Bolton and Ockenfels 2000), which makes the result interesting for applications beyond the family. In our framework, where social preferences are intention- rather than outcome-based, it turns out that mutual unkindness generally prevents equilibrium from being materially Pareto efficient, again in contrast to existing wisdom.

How, then, can the theoretical and empirical findings mentioned earlier be reconciled with the view presented in this paper? We will discuss this question in detail in Section 4. We will argue that, while it is possible to generate profitable gift-exchange in a variety of different theoretical models, this requires to make assumptions such as non-profit-maximization by the firm, or "extreme" reference points for fairness. Concerning the empirical findings, much of the laboratory evidence might in fact be best explained by two-sided social preferences. In contrast, some recent field experiments seem to provide support for the occurrence of negative rather than positive reciprocity in worker-firm relations (see Section 4). These findings are in line with our theoretical predictions.

The paper is organized as follows. In Section 2 we analyze a two-stage game as in Becker (1974) or Benjamin (2010), where the materialistic player moves first, followed by the reciprocal player. We derive the consequences of intention-based social preferences for

equilibrium kindness (2.2), the distribution of payoffs (2.3), equilibrium actions (2.4) and efficiency (2.5) in this framework.[4] In Section 3 we explore the implications of these results in two simple examples, a gift-exchange game and a moral hazard game. In Section 4 we discuss the relation of our results to several other papers, and Section 5 concludes. Some proofs and extensions can be found in the Appendix.

2 Model

2.1 Intentions in Action-Reaction Games

In this section, we investigate a class of two-stage "action-reaction" games with one materialistic and one reciprocal player. We will introduce two specific examples below. Our framework is also similar to the rotten kid setup (Benjamin 2010), except that we assume intention-based rather than outcome-based preferences.

There are two players. Player 1 moves first and chooses an action $w \in W$. The action becomes observable and player 2 reacts by choosing an action $e \in E$. Both W and E are compact subsets of $\mathbb{R}$, with minimal and maximal elements denoted by $\underline{w}$, $\overline{w}$ and $\underline{e}$, $\overline{e}$, respectively. In our examples, player 1 will be the *firm* that offers a *wage* and player 2 will be the *worker* who responds by supplying *effort*. In a rotten kid application, player 1 is interpreted as a child and player 2 as a parent.

A pure strategy of player 1 is simply an offer $a_1 \in A_1 = W$. A pure strategy of player 2 is a function $a_2 : W \to E$, from the set of strategies $A_2 = E^W$. We restrict attention to pure strategies throughout the body of the paper. With behavioral strategies, the additional question would arise how players attribute intentions to randomly determined outcomes.[5] The material payoffs of the two players as a function of the actions are denoted by $\tilde{\pi}_1(w, e)$ and $\tilde{\pi}_2(e, w)$. They are assumed to be continuous and differentiable on the interior of their domain. We can also obtain payoff functions defined on strategy profiles, by $\pi_1(a_1, a_2) = \tilde{\pi}_1(a_1, a_2(a_1))$ and $\pi_2(a_2, a_1) = \tilde{\pi}_2(a_2(a_1), a_1)$. An action profile (w, e) is *materially Pareto efficient* if there exists no other profile that strictly increases $\tilde{\pi}_i$ without reducing $\tilde{\pi}_j$, for $i, j \in \{1, 2\}$, $j \neq i$.

Player 1 is interested only in material payoffs. Player 2 has intention-based social prefer-

[4] Appendix A.1 contains an analysis of a more general dynamic framework. Considering dynamic interactions is important, as Fehr, Goette, and Zehnder (2009) have emphasized that reciprocity is magnified when firms and workers interact repeatedly. We can show that our negative kindness result holds for the general dynamic model, under very weak assumptions on the kindness norm.

[5] See Rabin (1993), Dufwenberg and Kirchsteiger (2004), Sobel (2005), Segal and Sobel (2007), Battigalli and Dufwenberg (2009) and Sebald (2010) for discussions of this issue. Our Appendix A.2 contains an analysis of a gift-exchange game with behavioral strategies, for the two most prominent approaches.

ences: She forms beliefs about the intention of player 1 and reacts correspondingly. We first specify how player 2 judges the kindness of her own behavior toward player 1. Conditional upon observing some w, we assume that responding by e generates a kindness of

$$\tilde{k}(e, w) = \tilde{\pi}_1(w, e) - \tilde{\pi}_1^e(w).$$

Intuitively, by choosing e player 2 gives a material payoff of $\tilde{\pi}_1(w, e)$ to player 1. This is compared to a reference standard, the *equitable payoff* $\tilde{\pi}_1^e(w)$, which measures what player 1 who has offered w deserves from the perspective of player 2. She believes to behave kindly ($\tilde{k} > 0$) when giving player 1 a payoff above $\tilde{\pi}_1^e(w)$ and unkindly ($\tilde{k} < 0$) when giving him a lower payoff.[6] Before we define the equitable payoff, we will specify how player 2 judges player 1's behavior, i.e. whether she attributes good or bad (or neutral) intentions to the offer of some w. We proceed analogously and compare the payoff that (player 2 believes that) player 1 believes to give to player 2 to an equitable payoff. Formally, we define

$$\tilde{\lambda}(w, c) = \tilde{\pi}_2(c(w), w) - \tilde{\pi}_2^e(c),$$

where $c \in A_2$ is player 2's second order belief. If player 1 offers w and player 2 behaves according to strategy c, then player 2 realizes a material payoff of $\tilde{\pi}_2(c(w), w)$. The use of c reflects the fact that player 1 chooses his action conditional on his first-order belief $b \in A_2$ about the strategy of player 2, who in turn forms a second-order belief c about this first-order belief, when trying to infer the intention behind an offer of w.

The equitable payoffs are determined as follows. Let $\Pi_1(w) = \{(\tilde{\pi}_1(w, e), \tilde{\pi}_2(e, w)) | e \in E\}$ be the set of payoff pairs that player 2 can induce if player 1 has offered w. Let $\Pi_1^E(w)$ be the set of Pareto efficient pairs in $\Pi_1(w)$, i.e. it contains those payoff pairs from $\Pi_1(w)$ for which there is no other pair in $\Pi_1(w)$ with a strictly larger payoff for one player and a payoff at least as large for the other player. Contrary to the concept of material Pareto efficiency introduced above, efficiency here is defined conditional on player 1's behavior. This approach follows Rabin (1993). Dufwenberg and Kirchsteiger (2004) proceed analogously but invoke a different notion of efficiency. This difference is important, and we will discuss it in detail in Section 4. Analogously, if player 2 follows strategy c, then player 1 can induce the payoff pairs in $\Pi_2(c) = \{(\tilde{\pi}_2(c(w), w), \tilde{\pi}_1(w, c(w))) | w \in W\}$. Let $\Pi_2^E(c)$ be the subset of Pareto efficient payoff pairs in the closure of $\Pi_2(c)$.[7] In our later examples, we will use the average

[6] In the generalized model introduced in Appendix A.1, the intended kindness of player i toward player j depends on i's strategy and i's updated *belief* about j's strategy. In the action-reaction game, since player 1 moves only once and the action becomes observable, we can identify player 2's updated belief with the observed w, so that the definition of $\tilde{k}$ does not have to rely on a first-order belief term explicitly.

[7] Since c is not necessarily continuous, it is possible that $\Pi_2(c)$ is not closed. Considering the closure

between player i's largest and smallest payoff within $\Pi_i^E(\cdot)$ as the equitable payoff. For our general results, however, we only impose the minimal requirement that $\tilde{\pi}_i^e(\cdot)$ does not correspond to an extreme point within this efficiency set.[8]

(A1) *(i)* If $|\Pi_i^E(\cdot)| \geq 2$, there exist $(\pi_i', \pi_j'), (\pi_i'', \pi_j'') \in \Pi_i^E(\cdot)$ with $\pi_i' < \tilde{\pi}_i^e(\cdot) < \pi_i''$.
(ii) If $\Pi_i^E(\cdot) = \{(\pi_i', \pi_j')\}$, then $\tilde{\pi}_i^e(\cdot) = \pi_i'$.

Assumption (A1) would, for example, allow the equitable payoff to depend on how costly it is for the opponent to give player i a larger payoff within the set Π_i^E.

To specify player 2's utility function, let $F : \mathbb{R}^2 \to \mathbb{R}$ be a function that assigns a psychological utility score $F(\tilde{k}, \tilde{\lambda})$ to each combination of kindness $\tilde{k}$ and belief about reciprocated kindness $\tilde{\lambda}$. We assume that $F(\tilde{k}, 0)$ is independent of $\tilde{k}$, i.e. 2's kindness has no impact on the own psychological utility if she expects to be treated neutrally. We also assume that $F(\tilde{k}, \tilde{\lambda})$ is strictly increasing in $\tilde{k}$ whenever $\tilde{\lambda} > 0$ and strictly decreasing if $\tilde{\lambda} < 0$, to capture the psychological utility gains from rewarding kindness with kindness and punishing unkindness with unkindness. We do not impose any additional assumptions on F. Rabin (1993) assumes $F(\tilde{k}, \tilde{\lambda}) = (1 + \tilde{k})\tilde{\lambda}$. In our later examples, we will follow Dufwenberg and Kirchsteiger (2004) and use $F(\tilde{k}, \tilde{\lambda}) = \tilde{k}\tilde{\lambda}$. Here, our more general formulation allows, for instance, for decreasing marginal psychological utility, i.e. the function F is not necessarily multilinear. Player 2's utility is now given by

$$\tilde{\pi}_2(e, w) + yF(\tilde{k}(e, w), \tilde{\lambda}(w, c)),$$

where $y > 0$ parameterizes the intensity of reciprocity.

Definition 1. *A strategy profile $(\hat{a}_1, \hat{a}_2)$ is an intentions-equilibrium (IE) if*
(i) $\hat{a}_1 \in \arg\max_{w \in W} \ \tilde{\pi}_1(w, \hat{a}_2(w))$, *and*
(ii) $\hat{a}_2(w) \in \arg\max_{e \in E} \ \tilde{\pi}_2(e, w) + yF(\tilde{k}(e, w), \tilde{\lambda}(w, \hat{a}_2))$ *for all* $w \in W$.

Intuitively, player 1 must maximize material payoffs in equilibrium, given his correct belief about player 2's reaction. Player 2 must maximize her utility, composed of material and psychological payoffs, in every subgame, i.e. for every conceivable offer $w \in W$. In evaluating the kindness of w, player 2 treats the correct second-order belief $c = \hat{a}_2$ as fixed.[9] Some of our results will compare the outcome of IE to the outcome of subgame-perfect equilibria

makes sure that $\Pi_2^E(c)$ is non-empty. This issue does not arise for $\Pi_1(w)$ above.

[8]Dufwenberg and Kirchsteiger (2004) work with the standard average specification. The appendix in Rabin (1993) contains a generalized model with assumptions analogous to (A1).

[9]This equilibrium definition is a special case of our more general treatment in Appendix A.1. The equilibrium is called intentions-equilibrium to distinguish our approach from the sequential reciprocity equilibrium of Dufwenberg and Kirchsteiger (2004).

when both players maximize their material payoffs. We will refer to these equilibria as *selfish equilibria* (SE). We conclude this subsection by introducing two simple examples that will be used to illustrate our general results.

Example 1: Gift-Exchange. In a simple gift-exchange game between a firm and a worker, we have $W = [0, \bar{w}]$ and $E = [0, \bar{e}]$. Material payoffs are given by $\tilde{\pi}_1(w, e) = e - w$ and $\tilde{\pi}_2(e, w) = v(w) - e$, for some continuously differentiable, strictly increasing and strictly concave function v that captures the worker's valuation of the wage. The wage is a gift because it has no direct incentivizing effect. Assuming an interior solution, the Pareto efficient action profile is characterized by the condition $v'(w^*) = 1$. We therefore speak of w^* as efficient wage; the choice of e does not affect material Pareto efficiency.[10] Denote the selfish equilibrium by $(\tilde{a}_1, \tilde{a}_2)$. We immediately obtain $\tilde{a}_2(w) = 0$ for all $w \in W$, which implies that there is a unique SE $(0, \tilde{a}_2)$ where no gift is given and no effort is provided. This inefficiency result has been the basis for much of the literature which investigates whether social preferences can help to overcome inefficiencies in employment relations.

Example 2: Moral Hazard. As a second example, consider a simple moral hazard variant. It will lead to additional insights on the relation between reciprocity, distribution and efficiency. Material payoffs are $\tilde{\pi}_1(w, e) = e(V - w)$ and $\tilde{\pi}_2(e, w) = ew - e^2$. Intuitively, the worker's effort e, which causes disutility e^2, is interpreted as the probability of completing a project successfully. The success of the project is observable, with payoff V in case of success and zero otherwise. The firm offers a bonus w to be paid (only) in case of success. We let $W = [0, V]$ and $E = [0, 1]$, and we assume that $V > 2$, which implies that full effort $e^* = 1$ is the materially Pareto efficient action.[11] It is straightforward to show that $\tilde{a}_2(w) = \min\{w/2, 1\}$ is the worker's material best response to wage w, and the firm will offer $\tilde{a}_1 = \min\{V/2, 2\}$ in the unique SE $(\tilde{a}_1, \tilde{a}_2)$. It is Pareto efficient if $V \geq 4$, but inefficient whenever $2 < V < 4$.

[10] We could just as well examine a game with a unique efficient effort level and purely redistributive wages. We have chosen our specification because it is directly comparable to the rotten kid framework, where the first mover's action determines efficiency.

[11] A binary version of the moral hazard game, with $E = \{0, p\}$, would be similar to the monopoly pricing game in Rabin (1993). The main differences are our dynamic structure, the fact that profits are decreasing in the wage as opposed to increasing in the monopoly price, and the assumption that only one player is reciprocal in our model. Rabin (1993) shows that positive kindness is impossible in the monopoly pricing game. This result is driven by the fact that the consumer can never be kind toward to monopolist, by construction of the game's payoffs.

2.2 Endogenous Bad Intentions

We first show that equilibrium behavior must be mutually unkind in any IE of the action-reaction game. This result is an important stepping stone in the analysis that will follow.

Proposition 1. *Suppose (A1) holds. Then, in any IE* $(\hat{a}_1, \hat{a}_2)$ *it holds that* $\tilde{k}(\hat{a}_2(\hat{a}_1), \hat{a}_1) \leq 0$*, with strict inequality if* $|\Pi_1^E(\hat{a}_1)| \geq 2$*, and* $\tilde{\lambda}(\hat{a}_1, \hat{a}_2) \leq 0$*, with strict inequality if* $|\Pi_2^E(\hat{a}_2)| \geq 2$*.*

Proof. Let $(\hat{a}_1, \hat{a}_2)$ be an IE. By definition we have $\hat{a}_1 \in \arg\max_{w \in W} \tilde{\pi}_1(w, \hat{a}_2(w))$, which in particular implies $\tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1)) \geq \pi_1$ for all $(\pi_2, \pi_1) \in \Pi_2^E(\hat{a}_2)$. By definition of Pareto efficiency of the payoff pairs in $\Pi_2^E(\hat{a}_2)$, we must therefore have $\tilde{\pi}_2(\hat{a}_2(\hat{a}_1), \hat{a}_1) \leq \pi_2$ for all $(\pi_2, \pi_1) \in \Pi_2^E(\hat{a}_2)$. Under assumption (A1) this implies $\tilde{\pi}_2(\hat{a}_2(\hat{a}_1), \hat{a}_1) \leq \tilde{\pi}_2^e(\hat{a}_2)$ and thus $\tilde{\lambda}(\hat{a}_1, \hat{a}_2) \leq 0$, with strict inequality if $|\Pi_2^E(\hat{a}_2)| \geq 2$. Next, again by definition of equilibrium, $\hat{a}_2(\hat{a}_1) \in \arg\max_{e \in E} \tilde{\pi}_2(e, \hat{a}_1) + yF(\tilde{k}(e, \hat{a}_1), \tilde{\lambda}(\hat{a}_1, \hat{a}_2))$ must hold. The second term is weakly decreasing in $\tilde{k}$ and hence in $\tilde{\pi}_1(\hat{a}_1, e)$, by the fact that $\tilde{\lambda}(\hat{a}_1, \hat{a}_2) \leq 0$. To show that $\tilde{k}(\hat{a}_2(\hat{a}_1), \hat{a}_1) \leq 0$, with strict inequality if $|\Pi_1^E(\hat{a}_1)| \geq 2$, we proceed by contradiction. Suppose first that $|\Pi_1^E(\hat{a}_1)| \geq 2$ but $\tilde{k}(\hat{a}_2(\hat{a}_1), \hat{a}_1) \geq 0$, i.e. $\tilde{\pi}_1^e(\hat{a}_1) \leq \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1))$. Then by (A1) there exists $(\pi_1', \pi_2') \in \Pi_1^E(\hat{a}_1)$ with $\pi_1' < \tilde{\pi}_1^e(\hat{a}_1) \leq \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1))$, and thus $\tilde{\pi}_2(\hat{a}_2(\hat{a}_1), \hat{a}_1) < \pi_2'$ by Pareto efficiency. This is a contradiction, as player 2 would strictly prefer the effort level that induces the payoff pair (π_1', π_2'). Suppose then that $\Pi_1^E(\hat{a}_1) = \{(\pi_1', \pi_2')\}$ but $\tilde{k}(\hat{a}_2(\hat{a}_1), \hat{a}_1) > 0$, i.e. $\pi_1' < \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1))$ under (A1). Pareto efficiency then again implies $\tilde{\pi}_2(\hat{a}_2(\hat{a}_1), \hat{a}_1) < \pi_2'$, a contradiction as before. □

Consider the perspective of player 1. Whenever there is a conflict of interest between the two players concerning the materially Pareto efficient allocations that player 1 can induce, he will avoid leaving more material payoff on the table than necessary, which is unkind behavior. Player 2 in turn does not expect to be treated kindly and does not benefit from being kind either. A kindness-neutral outcome is possible only if the two players' material interests coincide, so that payoff maximization by one player simultaneously maximizes the other player's payoff.[12] Importantly, this result does not depend on the particular two-stage model considered here. Appendix A.1 contains a generalized model of arbitrary finite-stage two-player games with observed actions and without moves of nature (which also includes normal-form games). We can show that negative kindness must prevail in equilibrium in every subgame that is reached on the equilibrium path, under an assumption analogous to (A1) and an additional mild assumption on the dynamic formation of equitable payoffs. In particular, the timing of the moves between the players is immaterial for this result to hold.

[12] Rabin (1993) has proven that an equilibrium with negative kindness always exists in his framework, among potentially other equilibria. We show that making one of the players materialistic eliminates any possibility for positive equilibrium kindness in our model, so that *only* unkindness survives.

The negative kindness result is in stark contrast to the rotten kid intuition outlined in the introduction. With suitable outcome-based social preferences on the side of the parent, a rotten kid anticipates monetary rewards for increasing family income. Maximizing joint income then becomes the kid's self-interest. But acts motivated by self-interest are not kind and would not be rewarded by a parent who cares about intentions, so that an analogous argument fails in our framework. As we will discuss in greater detail in Section 4, the resulting IE share quite a few properties with equilibria under envy preferences (Dur and Glazer 2008). This is remarkable as our model allows for both positive and negative emotions *a priori*, and envy-type behavior arises endogenously.

2.3 The Materialistic Player Suffers

To investigate the effect of reciprocity on the materialistic player's payoffs, we now impose an additional assumption that specifies the economic substance of the game.

(A2) (i) $\tilde{\pi}_1(w,e)$ is non-decreasing in e
(ii) For each $w \in W$, there is a unique e that maximizes $\tilde{\pi}_2(e,w)$ on E.

This assumption is satisfied by the two examples introduced earlier. While part (i) implies that player 1 weakly prefers larger values of e, (ii) allows for the possibility that player 2 finds some interior value of e desirable, as in the moral hazard example. For the rest of Section 2, we will assume that assumptions (A1) and (A2) are satisfied, without further mention. Player 2 then has a unique material best response $\tilde{a}_2(w) = \arg\max_{e\in E} \tilde{\pi}_2(e,w)$ to any action w. As an initial step, we are going to show under which conditions she deviates from this best response to punish player 1 for being unkind.

Lemma 1. *Any IE $(\hat{a}_1,\hat{a}_2)$ satisfies $\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1)$. The inequality is strict whenever $|\Pi_2^E(\hat{a}_2)| \geq 2$, $\tilde{a}_2(\hat{a}_1) \in \mathrm{int}\, E$, and $\tilde{\pi}_1(\hat{a}_1,e)$ is strictly increasing in e.*

Proof. Consider any IE $(\hat{a}_1,\hat{a}_2)$. By definition,

$$\hat{a}_2(\hat{a}_1) \in \arg\max_{e\in E} \tilde{\pi}_2(e,\hat{a}_1) + yF(\tilde{\pi}_1(\hat{a}_1,e) - \tilde{\pi}_1^e(\hat{a}_1),\tilde{\lambda}), \tag{1}$$

where $\tilde{\lambda} = \tilde{\lambda}(\hat{a}_1,\hat{a}_2)$ is independent of e and, according to Proposition 1, satisfies $\tilde{\lambda} \leq 0$, with strict inequality if $|\Pi_2^E(\hat{a}_2)| \geq 2$. Also by definition, $\tilde{a}_2(\hat{a}_1)$ is the unique maximizer of $\tilde{\pi}_2(e,\hat{a}_1)$, the first term in (1). Then, if $\tilde{\lambda} = 0$, we must have $\hat{a}_2(\hat{a}_1) = \tilde{a}_2(\hat{a}_1)$, because the second term in (1) is independent of e. If $\tilde{\lambda} < 0$, the reciprocity term $yF(\tilde{\pi}_1(\hat{a}_1,e) - \tilde{\pi}_1^e(\hat{a}_1),\tilde{\lambda})$ is strictly decreasing in $\tilde{\pi}_1(\hat{a}_1,e)$. Since $\tilde{\pi}_1(\hat{a}_1,e)$ is non-decreasing in e by (A2), we obtain

$\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1)$. If $\tilde{a}_2(\hat{a}_1) \in \operatorname{int} E$, then $\tilde{a}_2(\hat{a}_1)$ satisfies the necessary first order condition $\partial\tilde{\pi}_2(\tilde{a}_2(\hat{a}_1), \hat{a}_1)/\partial e = 0$. If, in addition, $\tilde{\pi}_1(\hat{a}_1, e)$ is strictly increasing in e, then the objective (1) is strictly decreasing in e at $e = \tilde{a}_2(\hat{a}_1)$, implying $\hat{a}_2(\hat{a}_1) < \tilde{a}_2(\hat{a}_1)$. □

The lemma states that player 2 reacts to the IE strategy $\hat{a}_1$ with a weakly smaller response than would be optimal from a purely materialistic perspective. The response is strictly lower when player 1 is strictly unkind in equilibrium ($|\Pi_2^E(\hat{a}_2)| \geq 2$), the materially optimal response is not a corner solution, and player 1 actually suffers strictly from a reduction in e. Based on this insight, we can now compare the equilibrium payoff of player 1 between the cases where player 2 is reciprocal (IE) and where she is materialistic (SE). Note that, although player 2 must clearly play strategy $\tilde{a}_2$ in any SE, multiple equilibria are still possible because player 1 could have more than one best reply $\tilde{a}_1$.

Proposition 2. *For any SE $(\tilde{a}_1, \tilde{a}_2)$ and any IE $(\hat{a}_1, \hat{a}_2)$ it holds that $\pi_1(\hat{a}_1, \hat{a}_2) \leq \pi_1(\tilde{a}_1, \tilde{a}_2)$. The inequality is strict whenever $|\Pi_2^E(\hat{a}_2)| \geq 2$, $\tilde{a}_2(\hat{a}_1) \in \operatorname{int} E$, and $\tilde{\pi}_1(\hat{a}_1, e)$ is strictly increasing in e.*

Proof. We have that $\pi_1(\hat{a}_1, \hat{a}_2) = \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1)) \leq \tilde{\pi}_1(\hat{a}_1, \tilde{a}_2(\hat{a}_1)) = \pi_1(\hat{a}_1, \tilde{a}_2)$, because $\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1)$ according to Lemma 1 and $\tilde{\pi}_1$ is non-decreasing in e. The inequality is strict if $|\Pi_2^E(\hat{a}_2)| \geq 2$, $\tilde{a}_2(\hat{a}_1) \in \operatorname{int} E$, and $\tilde{\pi}_1(\hat{a}_1, e)$ is strictly increasing in e, also according to Lemma 1. By contradiction, first assume $\pi_1(\tilde{a}_1, \tilde{a}_2) < \pi_1(\hat{a}_1, \hat{a}_2)$. Together with the above inequality this implies $\pi_1(\tilde{a}_1, \tilde{a}_2) < \pi_1(\hat{a}_1, \tilde{a}_2)$, which contradicts $\tilde{a}_1 \in \arg\max_{w \in W} \pi_1(w, \tilde{a}_2)$ and hence that $(\tilde{a}_1, \tilde{a}_2)$ is an SE. We obtain the analogous contradiction to the assumption that $\pi_1(\tilde{a}_1, \tilde{a}_2) \leq \pi_1(\hat{a}_1, \hat{a}_2)$ when the above inequality is strict. □

Proposition 2 shows that the materialistic player does not profit from facing an opponent who is reciprocal. His equilibrium payoff in *any* IE must necessarily be weakly smaller than in *any* SE, and strictly so whenever punishment actually takes place in the IE, which will be the case except if there are common interests or punishment is not viable. Note the key role of Proposition 1 for this result. The unkind behavior of player 1 leads to equilibrium punishment, which in turn reduces his payoffs. This result stands in contrast to results obtained for altruism in the rotten kid framework. For completeness, the following proposition confirms this claim with an altruism model that is similar in generality to our intention-based model. We assume that player 1 still maximizes material payoffs $\pi_1(a_1, a_2)$. Player 2 is altruistic, maximizing $\pi_2(a_2, a_1) + G(\pi_1(a_1, a_2))$, where G is an arbitrary but strictly increasing function of player 1's payoff. We are interested in subgame-perfect equilibria $(\bar{a}_1, \bar{a}_2)$ of the game with altruism, which we refer to as *altruism equilibria* (AE).

Proposition 3. *For any SE* $(\tilde{a}_1, \tilde{a}_2)$ *and any AE* $(\bar{a}_1, \bar{a}_2)$ *it holds that* $\pi_1(\tilde{a}_1, \tilde{a}_2) \leq \pi_1(\bar{a}_1, \bar{a}_2)$. *The inequality is strict whenever* $\tilde{a}_2(\tilde{a}_1) \in \operatorname{int} E$, *and* $\tilde{\pi}_1(\tilde{a}_1, e)$ *is strictly increasing in* e.

Proof. Arguing as in the proof of Lemma 1, it immediately follows that $\tilde{a}_2(\tilde{a}_1) \leq \bar{a}_2(\tilde{a}_1)$, with strict inequality if $\tilde{a}_2(\tilde{a}_1) \in \operatorname{int} E$ and $\tilde{\pi}_1(\tilde{a}_1, e)$ is strictly increasing in e. But then $\pi_1(\tilde{a}_1, \tilde{a}_2) = \tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(\tilde{a}_1, \bar{a}_2(\tilde{a}_1)) = \pi_1(\tilde{a}_1, \bar{a}_2)$, with strict inequality under the above conditions. Since $\bar{a}_1 \in \arg\max_{a_1 \in A_1} \pi_1(a_1, \bar{a}_2)$ by definition of AE, we obtain $\pi_1(\tilde{a}_1, \bar{a}_2) \leq \pi_1(\bar{a}_1, \bar{a}_2)$, which completes the proof. □

Propositions 2 and 3 together show that reciprocity and altruism have completely opposite effects concerning player 1's payoff. This can have quite important implications. For example, we could conclude that a profit-maximizing principal should try to hire altruistic agents because he can exploit their social preferences, while he should stay away from agents with intention-based social preferences as in our model.

2.4 Equilibrium Actions

Let us now examine how reciprocity affects the equilibrium actions. The upshot of the analysis will be that reciprocity typically has a positive effect on player 1's action, e.g. the equilibrium wage, while the effect on player 2's action, the effort, is more ambiguous. In this subsection, we will impose some additional but standard properties on the payoff functions, which are also satisfied by our previous examples.

(A3) (*i*) $\tilde{\pi}_1(w, e)$ is submodular on $W \times E$.
(*ii*) $\tilde{\pi}_2(e, w)$ is supermodular on $E \times W$.
(*iii*) $\tilde{\pi}_1(w, e)$ is weakly concave in e.

In particular, part (*ii*) of the assumption implies that player 2's material best response is weakly increasing in w. Part (*i*) requires the opposite for player 1, i.e. increasing w becomes weakly more costly for player 1 when e is larger.

For clarity, the following proposition applies to the simplified case when there is a unique SE $(\tilde{a}_1, \tilde{a}_2)$, i.e. a unique value of $w \in W$ that maximizes $\tilde{\pi}_1(w, \tilde{a}_2(w))$. The result is readily generalizable to allow for multiple SE, with its conclusion becoming a comparison between largest and/or smallest wages across equilibria.

Proposition 4. *Suppose (A3) holds and* $\tilde{\pi}_1(\tilde{a}_1, e)$ *is strictly increasing in* e. *Then* $\tilde{a}_1 \leq \hat{a}_1$ *holds for any IE* $(\hat{a}_1, \hat{a}_2)$ *in which* $\Delta(w) = \tilde{a}_2(w) - \hat{a}_2(w)$ *is weakly decreasing in* w *for all* $w \leq \tilde{a}_1$ *or for all* $w \geq \hat{a}_1$.

Proof. Step 1. First we show that $\hat{a}_2(\tilde{a}_1) \leq \tilde{a}_2(\tilde{a}_1)$ must hold. For a contradiction, assume $\hat{a}_2(\tilde{a}_1) > \tilde{a}_2(\tilde{a}_1)$. Then, $\pi_1(\tilde{a}_1, \tilde{a}_2) = \tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) < \tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1))$ under our assumptions. Since $\hat{a}_1 \in \arg\max_{w \in W} \tilde{\pi}_1(w, \hat{a}_2(w))$ by definition of IE, $\tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1)) = \pi_1(\hat{a}_1, \hat{a}_2)$ must hold, which implies $\pi_1(\tilde{a}_1, \tilde{a}_2) < \pi_1(\hat{a}_1, \hat{a}_2)$ and contradicts Proposition 2.

Step 2. To prove the proposition, we are going to show that, for any $w < \tilde{a}_1$, it holds that $\tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w, \tilde{a}_2(w)) \leq \tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w, \hat{a}_2(w))$. The LHS of this inequality is strictly positive by definition of $\tilde{a}_1$ as the unique maximizer of $\tilde{\pi}_1(w, \tilde{a}_2(w))$. Then, if the inequality holds, the RHS must also be strictly positive. But $\tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1))$ cannot be strictly positive, again because $\hat{a}_1$ maximizes $\tilde{\pi}_1(w, \hat{a}_2(w))$. Hence, once the inequality has been established, we know that $\hat{a}_1 \geq \tilde{a}_1$, which is the desired conclusion.

The inequality can be rearranged to $\tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w, \tilde{a}_2(w)) - \tilde{\pi}_1(w, \hat{a}_2(w))$. Now, $\tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w, \tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w, \hat{a}_2(\tilde{a}_1))$ holds due to $w < \tilde{a}_1$ and $\hat{a}_2(\tilde{a}_1) \leq \tilde{a}_2(\tilde{a}_1)$ from step 1, and submodularity of $\tilde{\pi}_1$ ((A3)(i)). Supermodularity of $\tilde{\pi}_2$ ((A3)(ii)) implies that $\tilde{a}_2(w) \leq \tilde{a}_2(\tilde{a}_1)$. Then, concavity of $\tilde{\pi}_1$ in e ((A3)(iii)) implies that $\tilde{\pi}_1(w, \tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w, \hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w, \tilde{a}_2(w)) - \tilde{\pi}_1(w, \hat{a}_2(\tilde{a}_1) - \tilde{a}_2(\tilde{a}_1) + \tilde{a}_2(w))$. Now observe that $\hat{a}_2(w) \leq \hat{a}_2(\tilde{a}_1) - \tilde{a}_2(\tilde{a}_1) + \tilde{a}_2(w)$, which follows immediately from the fact that $\Delta(w)$ is decreasing in w for $w \leq \tilde{a}_1$. Thus $\tilde{\pi}_1(w, \tilde{a}_2(w)) - \tilde{\pi}_1(w, \hat{a}_2(\tilde{a}_1) - \tilde{a}_2(\tilde{a}_1) + \tilde{a}_2(w)) \leq \tilde{\pi}_1(w, \tilde{a}_2(w)) - \tilde{\pi}_1(w, \hat{a}_2(w))$. Combining all these inequalities yields $\tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\tilde{a}_1, \hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w, \tilde{a}_2(w)) - \tilde{\pi}_1(w, \hat{a}_2(w))$, which is the desired result. The proof based on $\Delta(w)$ being decreasing for all $w \geq \hat{a}_1$ is analogous. □

The proposition applies to all those IE in which the punishment $\Delta(w)$ is weakly decreasing over a suitable range of wages. These are equilibria that preserve a certain property of the SE, namely that increasing w reduces unkindness when taking player 2's reaction into account, and hence leads player 2 to punish less. The proposition tells us that we can indeed expect reciprocity to have a positive impact on the level of a_1, e.g. on the firm's wage or on the beneficial action of a rotten kid, despite the fact that there is no benevolent gift-giving in our model. To grasp the intuition for the result, consider the case where $\Delta(w)$ is decreasing in w below $\tilde{a}_1$. There are now three effects of reciprocity that all tend to increase $\hat{a}_1$ over $\tilde{a}_1$. First, we already know that player 2 responds with lower e to $\hat{a}_1$ than materially optimal. Because player 1 is expecting this, he wants to choose a higher $\hat{a}_1$ by submodularity. The second effect arises from concavity of $\tilde{\pi}_1$: since e is lower, the marginal benefit of increasing $\hat{a}_1$ to induce a larger e is larger. Finally, the fact that $\Delta(w)$ is decreasing below $\tilde{a}_1$ implies that, for all a_1 smaller than the SE level $\tilde{a}_1$, the sensitivity of e with respect to a_1 is larger in the IE than in the SE. This again strengthens player 1's incentive to increase a_1.

This discussion also suggests an ambiguous effect on the equilibrium action of player 2. At the *margin*, inducing a larger value of e becomes easier for player 1, because player 2

reacts more strongly to higher wages for psychological reasons. Inducing a given *level* of e, however, is more costly for player 1, due to player 2's punishment, which by submodularity of player 1's objective means that the payoff increase resulting from a higher e is smaller.

2.5 Equilibrium Inefficiency

The previous results suggest that reciprocity can indeed work in favor of efficiency, for instance by increasing the level a_1 of a self-interested player's action beyond the inefficient SE level. We will now turn to the question of equilibrium efficiency more rigorously.

As player 2's utility is composed of material and psychological components, it becomes necessary to elaborate on the appropriate notion of efficiency. In his analysis of outcome-based social preferences, Benjamin (2010) distinguishes between *material* and *utility* efficiency. With intention-based preferences, the concept of overall utility efficiency is problematic. Knowing the outcome of an interaction is not sufficient to derive psychological utility, because it depends on the way the outcome was achieved. Hence a reasonable notion of utility efficiency might have to include the game, i.e. we might have to define a class of admissible games and define utility efficiency with respect to this class. Bierbrauer and Netzer (2012) follow this approach in a mechanism design framework with intention-based social preferences. Due to the complexity of the problem, here we refrain from analyzing utility efficiency and focus on material efficiency only.

Lemma 1 has shown that player 2's equilibrium action $\hat{a}_2(\hat{a}_1)$ in any IE is weakly and often strictly smaller than her material best-response $\tilde{a}_2(\hat{a}_1)$. Since player 1 suffers from the reduced response (at least weakly), this is equivalent to saying that player 2 punishes player 1 at an own material cost. Hence there is no hope to obtain an efficiency result in the spirit of the rotten kid theorem for intention-based social preferences. This is stated formally in the following proposition, which is an immediate corollary of Lemma 1.

Proposition 5. *Consider any IE* $(\hat{a}_1, \hat{a}_2)$ *in which* $|\Pi_2^E(\hat{a}_2)| \geq 2$, $\tilde{a}_2(\hat{a}_1) \in \operatorname{int} E$, *and* $\tilde{\pi}_1(\hat{a}_1, e)$ *is strictly increasing in* e. *Then* $(\hat{a}_1, \hat{a}_2)$ *is inefficient.*

3 Examples

3.1 Gift-Exchange

Consider the simple gift-exchange game introduced in Section 2.1, where $W = [0, \overline{w}]$ and $E = [0, \overline{e}]$. Since $\tilde{a}_2(w) = 0$ for all $w \in W$, it follows from Lemma 1 that $\hat{a}_2(\hat{a}_1) = 0$ must hold in any IE $(\hat{a}_1, \hat{a}_2)$. It then also follows immediately that $\hat{a}_1 = 0$ must be true, so that

intention-based social preferences do not help to solve the inefficiency problem.[13] As in the unique SE, both players choose zero values of their respective actions in any IE. Our analysis in Appendix A.2 reveals that this continues to hold when behavioral strategies are admitted, irrespective of their specific interpretation.

Rabin (1993, p. 1293f) presents an example of a normal-form gift-exchange game in which equilibria with mutual gift-giving exist.[14] This result is driven by the assumption that *both* the firm and the worker have intention-based social preferences, i.e. the firm's goal is not the maximization of profits. Falk and Fischbacher (2006, p. 305) also present a result according to which gift and effort are strictly positive even if only player 2 has social preferences.[15] In their model, outcome- and intention-based components are intertwined in the definition of social preferences. We return to this issue in the discussion section.

Notice that we do not make statements about unkindness and efforts in intentions-equilibria off the equilibrium path. Effort can be positive in IE for some non-equilibrium wages $w > 0$, because such wages might be interpreted as kind by the worker. Our argument is thus not driven by a general inability of the firm to exhibit kind behavior. It is driven by the fact that a kind wage will never induce enough effort to make it a profitable choice for the firm.

3.2 Moral Hazard

Now consider the moral hazard game from Section 2.1. To illustrate the effect that reciprocity has on equilibrium wages, we will first explore IE of a cut-off form, i.e. equilibria where

$$\hat{a}_2(w) = \begin{cases} 1 & \text{if} \quad \hat{w} \leq w, \\ 0 & \text{if} \quad w < \hat{w}, \end{cases}$$

for some $\hat{w} \in [0, V]$, and $\hat{a}_1 = \hat{w}$. We calculate equitable payoffs as the average between the largest and the smallest payoff in the efficiency set, and we use $F(\tilde{k}, \tilde{\lambda}) = \tilde{k}\tilde{\lambda}$.

Proposition 6. *Consider the moral hazard game.*
(i) IE of the cut-off form exist if and only if $y \geq 2/((V-1)^{3/2} - (V-1))$.
(ii) If the condition in (i) is satisfied, there exist values w^l *and* w^h *with* $2 < w^l \leq w^h < V$ *such that a cut-off profile is an IE if and only if* $\hat{w} \in [w^l, w^h]$.

[13] If $\hat{a}_1 > 0$ was true, we would have $\tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1)) = 0 - \hat{a}_1 < 0 \leq \hat{a}_2(0) - 0 = \tilde{\pi}_1(0, \hat{a}_2(0))$, a contradiction.

[14] Besides its normal-form, it also differs from the game considered here because effort is binary and the specification of material payoffs is different.

[15] Falk and Fischbacher (2006) use the payoff functions $\tilde{\pi}_1(w, e) = ve - w$ and $\tilde{\pi}_2(e, w) = w - \alpha e^2$. In their framework, Schliffke (2012) shows that an evolutionary process will yield efficiency, but with a strong payoff disadvantage for the firm.

Proof. See Appendix A.3 □

If cut-off equilibria exist, the set of possible equilibrium wages is characterized by lower and upper bounds w^l and w^h. The lower bound w^l is strictly larger than the wage $\tilde{a}_1$ that the firm would pay to an egoistic worker, and the reciprocal worker responds with maximal effort. This appears roughly in line with the theoretical and empirical literature discussed earlier. However, the large wage is still perceived as strictly unkind by the worker, which is an immediate corollary of Proposition 1. This contradicts the conventional interpretation of high wages as kind and low wages as unkind or selfish (see e.g. Card, DellaVigna, and Malmendier 2011, p. 50). Here, the large wage is paid out of purely egoistic considerations. The worker, in turn, requests this large wage because she would otherwise prefer to punish the firm's unkindness with reduced effort.

It can be shown that the bounds w^l and w^h exhibit straightforward comparative statics properties: increases in y and V both lead to increases of the lower and the upper bound. Intuitively, when y is small, rather low wages suffice to make the worker resist the temptation to punish. As psychological payoffs become more important, larger material payoffs and hence wages are required by the worker to still supply effort to the unkind firm. It is possible to show that $\lim_{y\to\infty} w^l = \lim_{y\to\infty} w^h = V$, i.e. in the limit the worker eventually reaps the complete gains from trade, although she cares less and less for material payoffs. Now consider the effect of V. The SE wage $\tilde{a}_1$ depends on V only to a limited extent. It is increasing in V for $V \leq 4$, when the firm wants to induce larger effort as V grows, but not beyond this point. The picture looks different with intention-based preferences, because the worker now also cares for the surplus left to the firm. For a given wage, increasing the project payoff V makes the option of punishment more attractive, because the worker can deprive the firm of higher profits by not supplying effort. A reciprocal player wants to be compensated for not using this sabotage option, reflected by smallest and largest equilibrium wages that are increasing in V throughout.[16] These findings could have implications for job design. On the one hand, we should expect jobs with more responsibility (as measured by V) to be paid better, even if they require exactly the same skills and efforts as lower responsibility jobs. On the other hand, employers could benefit from systematically structuring jobs so as to minimize potential for sabotage, even though sabotage is not observed in equilibrium.

IE with the cut-off property are materially Pareto efficient because they induce maximal effort. For $V \geq 4$ they share this property with the SE. If $V < 4$, however, the SE ceases to be

[16]There are counterparts to these comparative statics effects in Rabin's (1993) analysis of the monopoly pricing game. The monopolist's share of the gains from trade converges to one as production costs and the material value of the good increase to infinity. Also, the equilibrium price depends on production costs, which would not be the case with materialistic players.

efficient while cut-off IE continue to achieve the efficient outcome as long as V and y are not too low (so that the existence condition would be violated). In this case, one-sided intention-based social preferences help to solve efficiency problems in our moral hazard framework. Note, however, that this requires a relatively strong concern for reciprocity. At $V = 4$, for instance, the condition for existence of cut-off IE already requires $y \geq 2/(\sqrt{27} - 3) \approx 0.91$, so the worker has to put about equal weights on material and psychological payoffs.

The moral hazard game also has additional equilibria that are not of the cut-off form. The goal for the remainder of the section will be to illustrate one of these equilibria. We do not attempt a complete equilibrium analysis. Instead, we construct an equilibrium numerically.[17]

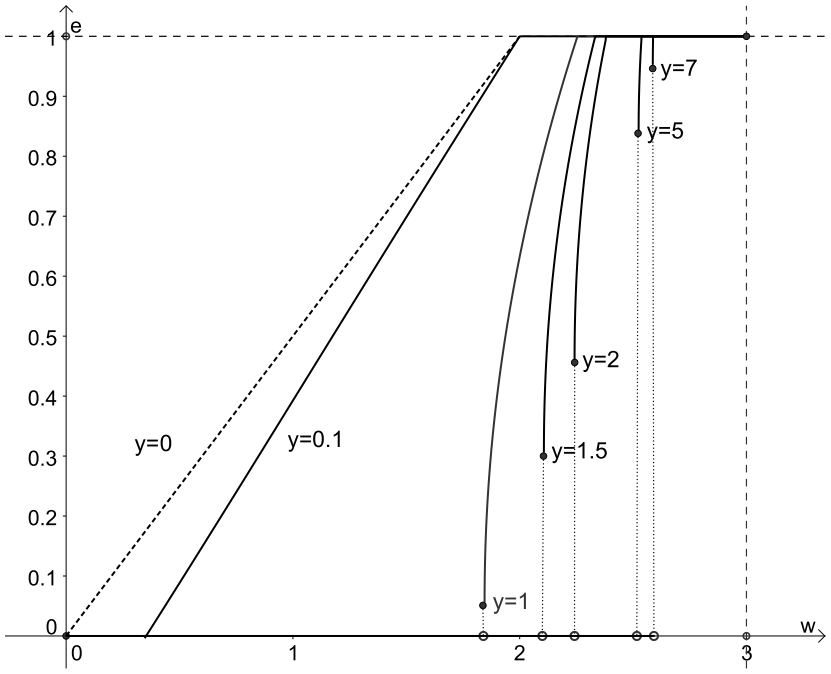

Figure 1: Equilibria for $V = 3$.

Figure 1 displays the worker's equilibrium strategy constructed in this way, for $V = 3$ and different values of y. The SE strategy $\tilde{a}_2$ is obtained for $y = 0$ and depicted as a dashed line. The firm offers the wage $\tilde{a}_1 = 3/2$ in the SE, inducing the inefficiently low effort level $\tilde{a}_2(\tilde{a}_1) = 3/4$. As we gradually increase the reciprocity weight y, the reaction function of the worker becomes flat up to some wage threshold, after which it starts to increase towards the maximal effort level. As we further increase y, a discontinuity emerges in the reaction

[17]We proceed as follows. First, given some second-order belief $c \in A_2$, the worker has a unique best effort response to any wage. The firm's preferred wage offer can then also be determined. Under the (potentially restrictive) assumption that the worker's material payoff is increasing in the wage in equilibrium, her equitable payoff can be calculated as the average between her equilibrium payoff and the payoff obtained for wage $w = V$. Then, the fixed point condition $c = \hat{a}_2$ can be invoked to determine the IE, which can finally be shown to satisfy the required monotone relation between wage and material payoff.

function, implying an upward jump from zero to strictly positive effort at some wage level. Note that the described strategy becomes increasingly similar to a cut-off strategy as y increases (for intermediate values of y, however, the equilibrium coexists with the previously described cut-off equilibria). Both the firm's equilibrium wage and the induced equilibrium effort level are monotone in y. This corroborates that reciprocity works in favor of material efficiency. However, full efficiency is only achieved for large values of y; the equilibrium wage induces less than the maximal effort when y is positive but small. Finally, the worker's material payoff increases and the firm's material payoff decreases in y.

4 Discussion

Our analysis reveals limitations of intention-based explanations for kind behavior such as gift-exchange between firms and workers. This of course begs the question what explains the empirically documented emergence of such phenomena. We will organize our discussion of this question along various theoretical models that generate gift-exchange equilibria. We attempt to assess the validity of these models as an explanation for the empirical findings. In doing so, we distinguish between evidence from the laboratory and from the field. We also delineate possible directions for future research.

4.1 Alternative Definitions of the Reference Point

Gift-exchange between a profit-maximizing firm and a reciprocal worker can be generated based on alternative definitions of the equitable payoff. The role of the reference point is in fact crucial for understanding a major difference between our contribution and other papers that model intentions in moral hazard related environments.[18] We have assumed that (i) the equitable payoff lies within the set of efficient payoffs as defined by Rabin (1993), and (ii) it is not an extreme point of this set. In contrast to (i), the concept of sequential reciprocity equilibrium due to Dufwenberg and Kirchsteiger (2004) requires to define $\Pi_i^E(\cdot)$ as the payoff pairs achievable when player $j \neq i$ can play any *efficient strategy* a_j (cf. the discussion in their Section 5). A strategy a_j is efficient except if there exists another strategy a'_j that always yields the same and sometimes higher payoffs to both players, where "always" refers to all histories and all strategies of player i. With this concept, the set $\Pi_i^E(\cdot)$ would potentially become substantially larger and include payoff pairs that are in fact Pareto dominated when

[18] Von Siemens (2009) and Dufwenberg, Smith, and Van Essen (2011) examine intention-based reciprocity in hold-up problems, and von Siemens (2013) analyzes the hidden cost of control (Falk and Kosfeld 2006). In the model of Hart and Moore (2008), contracts define reference points and might always leave one party ex-post angry and with a desire to retaliate.

player i's behavior is fixed. As a result, the reference point would not necessarily lie within the boundaries required by (A1), and equilibrium kindness would become possible even with one-sided reciprocity. Consider a cut-off equilibrium of the moral hazard game, for instance. Offering a wage below the agent's equilibrium cut-off value $\hat{w}$ would be efficient. Conditional on the agent's actual equilibrium strategy, the resulting outcome is Pareto inefficient. There are, however, non-equilibrium strategies of the agent for which the low wage would induce a Pareto efficient outcome, in which the principal would obtain very large payoffs. Hence the principal could be considered kind when offering $\hat{w}$, even though he holds a correct belief about the agent's actual equilibrium strategy and in fact does not sacrifice own payoffs in favor of the agent.[19] To see the role of (ii), it is useful to consider the related contribution of Englmaier and Leider (2012). The authors also model the interaction between a materialistic principal and a reciprocal agent. In their model, any wage above the conventional level, such as the market-clearing wage or the outside option, is assumed to be a benevolent gift and reciprocated by the worker. Hence the agent's equitable payoff can be thought of as an extreme point in the set of admissible payoff pairs. Similar mechanisms are also underlying the models in Akerlof (1982) and Bellemare and Shearer (2011). This leads to very different implications. Most importantly, the principal benefits from reciprocity. As soon as he offers more to the agent than the outside option, the agent responds by behaving kindly.[20]

While these alternative definitions of the reference point facilitate the emergence of gift-exchange equilibria, our approach clearly articulates an unease that seems to have worried some authors before. For instance, in his interpretation of experimental results, Charness (2004, p. 679) conjectures that employees might no longer perceive high wages as kind once they realize that paying these wages is in the employer's own interest. Fehr, Goette, and Zehnder (2009) emphasize the importance of explaining the fairness aspect of wage variations to the workers, and Bellemare and Shearer (2011) argue that gifts should not be "clearly in the short-term interests of the firm" (p. 861). Ultimately, however, the problem of the appropriate reference point model remains an important question for future empirical research.

[19]Based on this mechanism, in Sebald (2010) a profit-maximizing principal generates positive reciprocity by offering a high nominal wage that insures the agent against the risk of inflation, and in Dufwenberg and Kirchsteiger (2000) a profit-maximizing firm generates positive reciprocity by not hiring an outside worker at a lower wage. Another mechanism is proposed by von Siemens (2013), where a worker with privately known intention-based preferences reciprocates the behavior of a firm which is unkind towards herself, because the firm's behavior would have been kind towards a selfish worker.

[20]Englmaier and Leider (2012) address the question of whether efforts should be induced with relatively flat incentives, appealing to reciprocity, or with steep, strongly outcome-dependent incentives.

4.2 Two-Sided Concerns for Reciprocity

Equilibria with mutually kind behavior and profitable gift-exchange can also exist if social preferences are two-sided, i.e. if firms also have an intrinsic concern for their workers. Several intention-based models (e.g. Rabin 1993, Ruffle 1999, Arbak and Kranich 2005, Non 2012) rely on this assumption.[21]

In laboratory experiments (e.g. Fehr, Gächter and Kirchsteiger 1997, Charness 2004, Dhaene and Bouckaert 2010), the existence of two-sided social preferences appears plausible, as the roles of firms and workers are taken by the same student subjects. Even if laboratory subjects are good representatives of workers, however, a Lucas-type critique implies that their behavior cannot necessarily be extrapolated to real-world employment situations, where the firm might have different motives than the subject taking its role in the laboratory.[22] The intention-based model considered in this paper illustrates this problem transparently. Field studies are less susceptible to the critique, but the motives that workers presume behind experimental variations might still not coincide with those attributed to the same variations once they become part of the firm's ordinary compensation scheme. In Bellemare and Shearer (2011), for instance, the pay raise given to tree-planting workers is explained to them as resulting from the manager's deliberate choice to share a windfall gain (p. 858f). The workers' reaction might be different when the firm starts to exploit the identified behavioral correlation. Again, this poses interesting questions for future research about the attribution of motives to firms.[23]

[21] Rabin (1993) presents a labor contract example. Positive kindness is possible in this application because both employer and employee are assumed to have intention-based social preferences. Ruffle (1999) studies gift-giving, mostly focussing on surprise and related emotions in a non-strategic setting. In an extension to a simultaneous game with mutual reciprocity, he also obtains gift-exchange equilibria. Arbak and Kranich (2005) and Non (2012) are signalling models (Levine 1998) where preferences are assumed to be conditional on the type of the opponent (Gul and Pesendorfer 2010). In these models, the existence of some firms with social preferences might enable other firms to imitate and profit from strategic kindness. Such pooling equilibria can exist under specific circumstances in Arbak and Kranich (2005), but never in Non (2012).

[22] For related applications of the Lucas critique see Levitt and List (2007) and Bowles and Reyes (2009).

[23] Utikal and Fischbacher (2009) report on an experiment where individuals were asked to evaluate the intentions behind actions of a profit-maximizing firm: one treatment involved positive and one negative externalities on a third party. They find that, when the firm is in a dominant position and positive externalities are small, positive externalities are perceived as unintentional while negative externalities are perceived as fully intentional, broadly in line with our results. The effect is no longer present when the firm has small economic status, and the result is reversed if the positive externalities become sufficiently large. The approach of Utikal and Fischbacher (2009) is, however, not fully comparable to our model, for example because the person to judge and punish is not the one experiencing the externality.

4.3 Outcome-Based Preference Components

Gift-exchange can also arise in models of outcome-based social preferences. As discussed earlier, for the case of altruism this has already been recognized by Becker (1974, 1981), and for more general preferences including inequality-aversion it has been demonstrated by Benjamin (2010).[24]

Outcome-based preference components are most likely another part of the explanation of gift-exchange in the laboratory. The power of this approach to explain real-world gift-exchange between workers and firms is probably more limited. For inequality-aversion to generate gift-exchange, for instance, the firm's gift would have to be large enough to generate advantageous inequality for the worker (Card, DellaVigna, and Malmendier 2011, Kube, Maréchal and Puppe 2012b). Then, since intention- and outcome-based social preferences alone have trouble providing a convincing explanation, the same is presumably true for models which combine both motivations, but this is another interesting path for future research.[25] Dur and Glazer (2008) study optimal incentives when a worker envies the employer. While the employer is risk-neutral and has purely materialistic preferences, the agent is risk-averse and envious, with utility depending negatively on the difference between the principal's profit and wages. As in our model, profits decline as social preferences become more important. The bonus payment is positively affected by envy, and the effects on wages (base salaries) and efforts are ambiguous.[26] Hence our results for intention-based preferences are closer to those for models of envy than to those for models of altruism or inequity-aversion. This similarity arises endogenously in our framework, that a priori allows for both positive (altruism-type) or negative (envy-type) emotions.

4.4 Alternative Theories?

One could also think of theoretical explanations for gift-exchange beyond the class of social preference models discussed so far. For instance, social norms might prescribe certain customary reactions of workers to changes in the wage rate. Emotions such as gratitude or

[24] Itoh (2004), Englmaier and Wambach (2010), Bartling and von Siemens (2010) and Bartling (2011) are treatments of different moral hazard problems with inequality aversion and/or envy.

[25] Charness and Rabin (2002), Falk and Fischbacher (2006) and Cox, Friedman, and Gjerstad (2007) develop theories that account for both intention- and outcome-based social preferences. Evidence for the simultaneous presence of both types of preferences among laboratory subjects is provided by Andreoni, Brown and Vesterlund (2002), Charness and Haruvy (2002), and Falk, Fehr and Fischbacher (2003, 2008).

[26] Some of these results, such as declining profits, can also arise when the agent is inequality averse, provided she is worse off compared to the firm (see Itoh 2004). Dur and Glazer (2008) apply their model to argue that workers should be given stock options in spite of risk aversion, that stock options for the CEO have the additional cost that they increase worker envy, and to explain why the public sector (and non-profit organizations more generally) pay lower wages and use incentive pay less than the private sector.

feelings of obligation might also be relevant.

At the same time, several recent field studies (Gneezy and List 2006, Hennig-Schmidt, Rockenbach and Sadrieh 2010, Chemin, DeLaat and Kurmann 2011, Kube, Maréchal and Puppe 2012b) have found that, in the long run, negative reciprocity seems to be more robust than positive reciprocity.[27] This is consistent with our view that profitable gift-exchange should cease to exist as soon as workers are fully aware that the firm's ultimate goal is profit-maximization. The fact that the laboratory findings are not too puzzling from our theory's perspective, together with the indication of considerably less gift-exchange in the field, suggests that a search for new theories might be less urgent than it appears on first glance. More empirical work about real-world labor relations will be necessary, however, before definite conclusions along these lines can be drawn.

5 Conclusions

We have explored the limitations of intention-based social preferences as an explanation for profitable gift-exchange, by considering the interaction between a self-interested and a reciprocal player. In a gift-exchange game, the equilibrium never involves gift-giving. In a moral hazard game, reciprocity can affect behavior in a way that superficially looks as though it involves positive kindness, because wages and effort are higher compared to the benchmark of two self-interested players. However, a careful analysis shows that this behavior does not correspond to positive reciprocity. The agent still perceives the received wage as less than equitable and punishes the principal accordingly. Efforts are not high enough to compensate for higher wages, and as a result the principal obtains lower payoffs than when he faces a self-interested agent. Thus, even though reciprocity has important implications for behavior and also efficiency, it cannot be exploited by a self-interested player.

Our analysis does not question the role of positive kindness in the lab, where two-sided reciprocity is likely to be important. However, it casts doubt on the applicability of similar reasoning in the field, where there is less symmetry between players. Our analysis is thus consistent with recent evidence suggesting that positive reciprocity is less common in the field than in the lab.

Suitable extensions of our model could be used to obtain results on organizational design (Englmaier and Leider 2012, von Siemens 2011). For instance, experimental observations

[27] In the field study by Kube, Maréchal, and Puppe (2012a), non-monetary gifts induce positive reciprocity, while simple monetary gifts do not. Such behavior could arise in the model of Dur (2009), provided that non-monetary gifts convey attention and are used for signalling by an altruistic principal. Dur (2009) argues that monetary gift-exchange might work in the laboratory but not in the field because principals prefer giving non-monetary gifts in real-world employment relations.

suggest that, if a principal gives the control rights for unpopular decisions to third parties, he may benefit because he is perceived as less unkind than when he takes such decisions himself (Bartling and Fischbacher 2012). It would be interesting to see whether such behavior is consistent with our framework. Finally, extensions would seem suitable to shed new light on the longstanding debate on the boundaries of the firm. Nickerson and Zenger (2008) argue that larger firms might suffer from increased costs due to social comparisons between employees. From our perspective, changes in the numbers of employees working on related projects may affect their potential for sabotage and thus the potential adverse consequences of reciprocal behavior.

References

Akerlof, G. (1982): "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, 97, 543–569.

Andreoni, J., P. Brown, and L. Vesterlund (2002): "What Makes an Allocation Fair? Some Experimental Evidence," *Games and Economic Behavior*, 40, 1–24.

Arbak, E., and L. Kranich (2005): "Can Wages Signal Kindness," Working paper, CNRS.

Bartling, B. (2011): "Relative Performance or Team Evaluation? Optimal Contracts for Other-Regarding Agents," *Journal of Economic Behavior and Organization*, 79, 183–193.

Bartling, B., and U. Fischbacher (2012): "Shifting the Blame: On Delegation and Responsibility," *Review of Economic Studies*, 79, 67–87.

Bartling, B., and F. von Siemens (2010): "The Intensity of Incentives in Firms and Markets: Moral Hazard With Envious Agents," *Labour Economics*, 17, 598–607.

Battigalli, P., and M. Dufwenberg (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144, 1–35.

Becker, G. (1974): "A Theory of Social Interactions," *Journal of Political Economy*, 82, 1063–1093.

——— (1981): *A Treatise on the Family*. Harvard University Press, Cambridge, Massachusetts.

Bellemare, C., and B. Shearer (2011): "On the Relevance and Composition of Gifts Within the Firm: Evidence from Field Experiments," *International Economic Review*, 52, 855–882.

Benjamin, D. (2010): "Social Preferences and the Efficiency of Bilateral Exchange," mimeo, Cornell University.

Bierbrauer, F., and N. Netzer (2012): "Mechanism Design and Intentions," Discussion paper, UZH Department of Economics, No. 66.

Bolton, G., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Bowles, S., and S. Reyes (2009): "Economic Incentives and Social Preferences: A Preference-Based Lucas Critique of Public Policy," CESifo Working Paper No. 2734.

Card, D., S. DellaVigna, and U. Malmendier (2011): "The Role of Theory in Field Experiments," *Journal of Economic Perspectives*, 25, 39–62.

Charness, G. (2004): "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labour Economics*, 22, 665–688.

Charness, G., and E. Haruvy (2002): "Altruism, Equity, and Reciprocity in a Gift-Exchange Experiment: An Encompassing Approach," *Games and Economic Behavior*, 40, 203–231.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817–869.

Chemin, M., J. DeLaat, and A. Kurmann (2011): "Reciprocity in Labor Relations: Evidence from a Field Experiment with Long-Term Relationships," Working Paper 11-27, CIRPEE.

Cox, J., D. Friedman, and S. Gjerstad (2007): "A Tractable Model of Reciprocity and Fairness," *Games and Economic Behavior*, 59, 17–45.

Dhaene, G., and J. Bouckaert (2010): "Sequential Reciprocity in Two-Player, Two-Stage Games: An Experimental Analysis," *Games and Economic Behavior*, 70, 289–303.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2009): "Homo Reciprocans: Survey Evidence on Behavioral Outcomes," *Economic Journal*, 119, 592–612.

Dufwenberg, M., and G. Kirchsteiger (2000): "Reciprocity and Wage Undercutting," *European Economic Review*, 44, 1069–1078.

——— (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268–298.

Dufwenberg, M., A. Smith, and M. Van Essen (2011): "Hold-up: With a Vengeance," *Economic Inquiry*, pp. 1–13.

Dur, R. (2009): "Gift Exchange in the Workplace: Money or Attention?," *Journal of the European Economic Association*, 7, 550–560.

Dur, R., and A. Glazer (2008): "Optimal Contracts When a Worker Envies His Boss," *Journal of Law, Economics, and Organization*, 24, 120–137.

Englmaier, F., and S. Leider (2012): "Contractual and Organizational Structure with Reciprocal Agents," *American Economic Journal: Microeconomics*, forthcoming.

Englmaier, F., and A. Wambach (2010): "Optimal Incentive Contracts Under Inequity Aversion," *Games and Economic Behavior*, 69, 312–328.

Falk, A., E. Fehr, and U. Fischbacher (2003): "On the Nature of Fair Behavior," *Economic Inquiry*, 41, 20–26.

——— (2008): "Testing Theories of Fairness - Intentions Matter," *Games and Economic Behavior*, 62, 287–303.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54, 293–315.

Falk, A., and M. Kosfeld (2006): "The Hidden Cost of Control," *American Economic Review*, 96, 1611–1630.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica*, 65, 833–860.

Fehr, E., L. Goette, and C. Zehnder (2009): "A Behavioral Account of the Labor Market: The Role of Fairness Concerns," *Annual Review of Economics*, 1, 355–384.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, 108, 437–459.

Fehr, E., and K. Schmidt (1999): "A Theory Of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological games and sequential rationality," *Games and Economic Behavior*, 1, 60–79.

Gneezy, U., and J. List (2006): "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, 74, 1365–1384.

Gul, F., and W. Pesendorfer (2010): "Interdependent Preference Models as a Theory of Intentions," mimeo.

Hart, O., and J. Moore (2008): "Contracts as Reference Points," *Quarterly Journal of Economics*, 123, 1–48.

Hennig-Schmidt, H., B. Rockenbach, and A. Sadrieh (2010): "In Search of Workers' Real Effort Reciprocity – a Field and a Laboratory Experiment," *Journal of the European Economic Association*, 8, 817–837.

Itoh, H. (2004): "Moral Hazard and Other-Regarding Preferences," *Japanese Economic Review*, 55, 18–45.

Kube, S., M. Maréchal, and C. Puppe (2012a): "The Currency of Reciprocity - Gift-Exchange in the Workplace," *American Economic Review*, 102, 1644–1662.

——— (2012b): "Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment," *Journal of the European Economic Association*, forthcoming.

Levine, D. (1998): "Modelling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1, 593–622.

Levitt, S., and J. List (2007): "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 21, 153–174.

McCabe, K., M. Rigdon, and V. Smith (2003): "Positive Reciprocity and Intentions in Trust Games," *Journal of Economic Behavior and Organization*, 52, 267–275.

Nickerson, J., and T. Zenger (2008): "Envy, Comparison Costs, and the Economic Theory of the Firm," *Strategic Management Journal*, 29, 1429–1449.

Non, A. (2012): "Gift-Exchange, Incentives, and Heterogeneous Workers," *Games and Economic Behavior*, 75, 319–336.

Offerman, T. (2002): "Hurting Hurts More Than Helping Helps," *European Economic Review*, 46, 1423–1437.

Rabin, M. (1993): "Incorporating Fairness Into Game Theory and Economics," *American Economic Review*, 83, 1281–1302.

Ruffle, B. (1999): "Gift Giving With Emotions," *Journal of Economic Behavior and Organization*, 39, 399–420.

Schliffke, P. (2012): "The Co-Evolution of Reciprocity-Based Wage Offers and Effort Choices," *Economics Letters*, 117, 326–329.

Sebald, A. (2010): "Attribution and Reciprocity," *Games and Economic Behavior*, 68, 339–352.

Segal, U., and J. Sobel (2007): "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," *Journal of Economic Theory*, 136, 197–216.

Sobel, J. (2005): "Interdependent Preferences and Reciprocity," *Journal of Economic Literature*, 43, 392–436.

Utikal, V., and U. Fischbacher (2009): "On the Attribution of Externalities," Discussion paper, TWI, No. 46.

von Siemens, F. (2009): "Bargaining under Incomplete Information, Fairness, and the Hold-Up Problem," *Journal of Economic Behavior and Organization*, 71, 486–494.

——— (2011): "Heterogeneous Social Preferences, Screening, and Employment Contracts," *Oxford Economic Papers*, 63, 499–522.

——— (2013): "Intention-Based Reciprocity and the Hidden Costs of Control," *Journal of Economic Behavior and Organization*, 92, 55–65.

A Appendix

A.1 Generalized Model

In this appendix, we introduce a framework that generalizes the model considered in the body of the text, by allowing for very general dynamic interactions. We closely follow Dufwenberg and Kirchsteiger (2004), with some differences to be emphasized below.

Game and Strategies. We consider a two-player game with finitely many stages, where the players interact simultaneously every period and the outcome of the interaction becomes observable before the next period. Sequential moves of the players can be captured by assigning singleton action sets to players in periods where they are inactive. H denotes the set of histories, where each history is a list of previous actions profiles. The symbol $\varnothing \in H$ represents the root of the game. We let A_i be the set of pure strategies for player i, with elements $a_i \in A_i$ that are mappings from H to available actions at the corresponding information set. Given $a_i \in A_i$ and $h \in H$, we denote by a_i^h

the strategy obtained from a_i by replacing all actions that are inconsistent with h by those which are consistent with h, and keeping a_i unchanged otherwise. Hence a_i^h is an updated strategy that coincides with a_i except that, everywhere along the history h, it prescribes the action for player i that correspond to the subsequent element in h. In particular, $a_i^{\varnothing} = a_i$.[28]

Beliefs. To be able to model the intentions that players attribute to each other's behavior, we need to introduce beliefs about strategies and beliefs. Using the notation of Rabin (1993) and Dufwenberg and Kirchsteiger (2004), we denote first-order point beliefs by the letter b, i.e. $b_{ij} \in A_j$ is the belief of player i about the strategy a_j of player j, for $i = 1, 2$ and $j \neq i$. Similarly, second-order point beliefs are denoted by the letter c, so that $c_{iji} \in A_i$ is the belief of player i about the first-order belief b_{ji} of player j (about the strategy a_i of i). Beliefs can be updated exactly like strategies. Given b_{ij} and a history h, let b_{ij}^h be the new belief updated from b_{ij}, by replacing all actions that are inconsistent with h by those which are consistent. Analogously, c_{iji}^h is the second order belief updated from c_{iji}, using player i's actions that lead to h. We again have $b_{ij}^{\varnothing} = b_{ij}$ and $c_{iji}^{\varnothing} = c_{iji}$. All beliefs will be required to be correct in equilibrium.

Material Payoffs. In line with the earlier notation, let $\tilde{\pi}_i$ denote player i's material payoff function defined on the set of terminal nodes of the game. We assume that the set of possible payoff pairs is bounded. Based on $\tilde{\pi}_i$ we can also define material payoffs π_i on the set of strategy profiles $A = A_1 \times A_2$, so that $\pi_i(a_i, a_j)$ is player i's material payoff in the terminal node reached under (a_i, a_j). For any history $h \in H$, let $\pi_i(a_i, a_j|h) = \pi_i(a_i^h, a_j^h)$ denote the payoff of player i if the updated profile (a_i^h, a_j^h) is played instead of (a_i, a_j).

Kindness. Based on any collection $(a_i, b_{ij}, c_{iji})_{i,j=1,2,\, i\neq j}$, we can now assign measures of kindness and beliefs about them to every information set $h \in H$. We denote by $k_{ij}(a_i, b_{ij}|h)$ the kindness of i to j in information set h.[29] We let $\lambda_{iji}(b_{ij}, c_{iji}|h)$ denote player i's belief about how kind j is to him in information set h. For any $h \in H$, the equitable payoff for player i in this information set is denoted by $\pi_i^e(a_i|h)$. Now, the kindness terms are

$$k_{ij}(a_i, b_{ij}|h) = \pi_j(b_{ij}, a_i|h) - \pi_j^e(b_{ij}|h)$$

and

$$\lambda_{iji}(b_{ij}, c_{iji}|h) = \pi_i(c_{iji}, b_{ij}|h) - \pi_i^e(c_{iji}|h).$$

To specify the equitable payoffs, let $\Pi_i(a_i|h) = \{(\pi_i(a_i^h, a_j), \pi_j(a_j, a_i^h))|a_j \in A_j\}$ be the set of payoff pairs that can be achieved if player i plays strategy a_i^h while player j plays an arbitrary strategy. Let $\Pi_i^E(a_i|h)$ be the Pareto efficient payoff pairs in the closure of $\Pi_i(a_i|h)$. The following two assumptions describe minimal requirements that equitable payoffs have to satisfy for every player, strategy, and history.

[28] Dufwenberg and Kirchsteiger (2004) consider n players but restrict attention to finite action sets.

[29] More accurately, k_{ij} should be called player i's belief about how kind he is to player j, because it is based on the belief b_{ij}. We refrain from using this formulation for simplicity.

(A1') (*i*) If $|\Pi_i^E(a_i|h)| \geq 2$, there exist $(\pi_i', \pi_j'), (\pi_i'', \pi_j'') \in \Pi_i^E(a_i|h)$ with $\pi_i' < \pi_i^e(a_i|h) < \pi_i''$.
(*ii*) If $\Pi_i^E(a_i|h) = \{(\pi_i', \pi_j')\}$, then $\pi_i^e(a_i|h) = \pi_i'$.

Assumption (A1') corresponds to (A1) in the body of the text, and requires that the equitable payoff is not extreme in the set of Pareto efficient payoff pairs. This is obviously fulfilled by the standard average specification. Another obvious property of the standard specification is that the equitable payoff depends only on the set of achievable payoffs for the player under consideration. We only need two implications of this property for our analysis, which are formulated in the following assumption.[30]

(A2') (*i*) $\pi_i^e(a_i|h) = \pi_i^e(a_i^h|h)$.
(*ii*) $\pi_i^e(a_i|h) = \pi_i^e(a_i|h')$ if $a_i^h = a_i^{h'}$.

Assumption (A2')(*i*) ensures that, for all strategies which coincide except possibly along history h, the equitable payoff in h will be the same. As soon as we have reached h and updated accordingly, these strategies have become identical, and so have the sets of potentially achievable payoffs. Hence (A2')(*i*) rules out that the equitable payoff in h varies with the deviation that was necessary to reach h. Part (*ii*) ensures that if a player has reached a history h, possibly by deviating from a_i, but then continues to play according to a_i up to some later history h', then his equitable payoff will be the same in h and h'. In particular, this also ensures that the equitable payoffs do not change along the equilibrium path.

Utility. With a function F as described before, for every information set $h \in H$ and player i let

$$U_i(a_i, b_{ij}, c_{iji}|h) = \pi_i(a_i, b_{ij}|h) + y_i F\left(k_{ij}(a_i, b_{ij}|h), \lambda_{iji}(b_{ij}, c_{iji}|h)\right) \tag{2}$$

be player i's utility in information set h, which is based on material payoffs from updated strategies and contains the updated reciprocity term added with a weight of $y_i \geq 0$.

Equilibrium. Following Dufwenberg and Kirchsteiger (2004), we require sequential rationality, that is, each player must maximize U_i in each information set $h \in H$.[31]

Definition 2. *A strategy profile $(\hat{a}_1, \hat{a}_2) \in A$ is an intentions-equilibrium (IE) if, for both $i, j \in \{1, 2\}$, $j \neq i$, and all $h \in H$,*
(i) $\hat{a}_i \in \arg\max_{a_i \in A_i} U_i(a_i, b_{ij}, c_{iji}|h)$,
(ii) $b_{ij} = \hat{a}_j$, *and*
(iii) $c_{iji} = \hat{a}_i$.

[30] While Rabin (1993) provides an assumption analogous to (A1'), our assumption (A2') is particular to the dynamic setup considered here.

[31] Dufwenberg and Kirchsteiger (2004) rule out profitable one-shot deviations only, while we preclude the existence of any arbitrary profitable deviation. This makes a difference with intention-based preferences, which can be dynamically inconsistent (see p. 279 in Dufwenberg and Kirchsteiger (2004) for a discussion). An equilibrium in our sense survives even if players can commit to multi-stage deviations, or if they lack the sophistication to understand that a seemingly profitable multi-stage deviation might no longer appear profitable later on in the game tree. In our two-stage model, the difference becomes irrelevant.

The equilibrium definition in Section 2.1 is a special case of this general formulation. If player 1 is selfish and moves only once, at the root of the game, his optimality condition boils down to maximization of material payoffs $\pi_1(a_1, \hat{a}_2) = \tilde{\pi}_1(w, \hat{a}_2(w))$ by choice of an offer w, given correct equilibrium beliefs about player 2's strategy. Each possible offer $w \in W$ constitutes a history, so player 2 must maximize utility for every w, given correct and updated equilibrium beliefs. Observe that updating to history $h = w$ yields $a_2^h = a_2$, $b_{21}^h = w$ and $c_{212}^h = c_{212}$. Substituting this into player 2's utility yields the simplified expression in Section 2.1, where Assumption (A2') is used to simplify $\pi_1^e(b_{21}|w) = \pi_1^e(b_{21}^w|w) = \pi_1^e(w|w) =: \tilde{\pi}_1^e(w)$ and $\pi_2^e(c_{212}|w) = \pi_2^e(c_{212}|\varnothing) =: \tilde{\pi}_2^e(c_{212})$.

Result. If one of the players is selfish (say player 1, without loss of generality), the observation that equilibrium behavior must be mutually unkind holds in our fully general framework.

Proposition 7. *Suppose (A1') and (A2') hold and $y_1 = 0$, $y_2 \geq 0$. Then, in any IE $(\hat{a}_1, \hat{a}_2)$ it holds that $k_{ij}(\hat{a}_i, \hat{a}_j|h) \leq 0$ for both $i, j \in \{1, 2\}, i \neq j$, and any $h \in H$ that is reached on the equilibrium path. The inequality for k_{ij} is strict if $|\Pi_j^E(\hat{a}_j|\varnothing)| \geq 2$.*

Proof. Step 1. Let $(\hat{a}_1, \hat{a}_2)$ be an IE, and first consider history $h = \varnothing$. Since $y_1 = 0$, player 1's optimality condition in $h = \varnothing$ can be written as $\hat{a}_1 \in \arg\max_{a_1 \in A_1} \pi_1(a_1, \hat{a}_2^{\varnothing})$. The fact that, under assumption (A1'), it holds that $k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) \leq 0$, with strict inequality if $|\Pi_2^E(\hat{a}_2|\varnothing)| \geq 2$, now follows exactly as in the proof of Proposition 1. Consider then player 2, whose optimality condition in $h = \varnothing$ can be written as $\hat{a}_2 \in \arg\max_{a_2 \in A_2} \pi_2(a_2, \hat{a}_1^{\varnothing}) + y_2 F(k_{21}(a_2, \hat{a}_1|\varnothing), \lambda_{212}(\hat{a}_1, \hat{a}_2|\varnothing))$. Since $\lambda_{212}(\hat{a}_1, \hat{a}_2|\varnothing) = k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) \leq 0$, player 2's utility is weakly decreasing in $k_{21}(a_2, \hat{a}_1|\varnothing)$ and hence in $\pi_1(\hat{a}_1^{\varnothing}, a_2)$. The fact that $k_{21}(\hat{a}_2, \hat{a}_1|\varnothing) \leq 0$, with strict inequality if $|\Pi_1^E(\hat{a}_1|\varnothing)| \geq 2$, now again follows as in the proof of Proposition 1, with the additional argument that payoff pairs arbitrarily close to (π_1', π_2') can be induced by a profitable deviation of player 2 even if $(\pi_1', \pi_2') \in \Pi_1^E(\hat{a}_1) \backslash \Pi_1(\hat{a}_1)$.

Step 2. Consider now any history h on the equilibrium path, i.e. that satisfies $\hat{a}_i^h = \hat{a}_i$ for $i = 1, 2$. We then have that $\pi_i(\hat{a}_i, \hat{a}_j|h) = \pi_i(\hat{a}_i, \hat{a}_j|\varnothing)$ and $\pi_i^e(\hat{a}_i|h) = \pi_i^e(\hat{a}_i|\varnothing)$ for $i = 1, 2$, $i \neq j$, the latter by (A2'). Hence $k_{ij}(\hat{a}_i, \hat{a}_j|h) = k_{ij}(\hat{a}_i, \hat{a}_j|\varnothing)$ anywhere on the equilibrium path. □

A.2 Gift-Exchange with Behavioral Strategies

In this appendix, we analyze the gift-exchange game from Section 2.1 when behavioral strategies are allowed. As Rabin (1993) has already pointed out, the interpretation of a behavioral strategy becomes important with intention-based social preferences. We will consider two different approaches. First, we follow the mass-action interpretation of Dufwenberg and Kirchsteiger (2004), according to which behavioral strategies reflect population frequencies of pure strategies. Consequently, the outcome of a player's randomization will be interpreted as fully intentional by the opponent. We refer to this approach as "implicit randomization". Alternatively, a player might randomize consciously. The intention that the opponent attributes to such behavior will not depend on the realized outcome of the randomization. Sebald (2010) has provided a formal framework that allows to model this interpretation, which we refer to as "explicit randomization".

A.2.1 Implicit Randomization

In contrast to Section 2.1, let a_1 be a behavioral strategy of the firm, which is a probability measure on (the Borel σ-algebra of) W. Let $A_1 = \Delta W$ be the set of all such probability measures. We will continue to write $a_1 = w$ for the Dirac measure for w, i.e. the measure that corresponds to pure strategy w. Analogously, a_2 is a behavioral strategy of the worker, which assigns a probability measure $a_2(w) \in \Delta E$ to every history $w \in W$, where ΔE denotes the set of all probability measures on E. We again write $a_2(w) = e$ for Dirac measures. Let $A_2 = (\Delta E)^W$ be the set of all behavioral strategies of the worker. The material payoffs $\tilde{\pi}_1(w, e)$ and $\tilde{\pi}_2(e, w)$ at terminal nodes of the game are as defined in Section 2.1. It is convenient to also write $\tilde{\pi}_1(w, \check{e})$ and $\tilde{\pi}_2(\check{e}, w)$ for the expected payoffs if the firm pays a wage of w and the worker reacts according to $\check{e} \in \Delta E$. Moreover, we can derive the expected payoffs $\pi_1(a_1, a_2)$ and $\pi_2(a_2, a_1)$ for behavioral strategy profiles, based on the induced probability measure over terminal nodes.

The definition of the worker's kindness toward the firm is a straightforward generalization of the concept from Section 2.1. Given an arbitrary history $w \in W$, the worker can induce the expected payoff pairs $\Pi_1(w) = \{(\tilde{\pi}_1(w, \check{e}), \tilde{\pi}_2(\check{e}, w)) | \check{e} \in \Delta E\}$, and the equitable payoff $\tilde{\pi}_1^e(w)$ for the firm in history w is defined based on the Pareto efficient pairs from $\Pi_1(w)$ according to assumption (A1). Kindness is then given by

$$\tilde{k}(\check{e}, w) = \tilde{\pi}_1(w, \check{e}) - \tilde{\pi}_1^e(w).$$

Suppose the firm plays a_1 and consider some history w. In the model of Dufwenberg and Kirchsteiger (2004), the worker now attributes the same intention to the firm as if the latter had played the pure strategy w from the outset. Hence the worker forms the belief

$$\tilde{\lambda}^{IR}(w, c) = \tilde{\pi}_2(c(w), w) - \tilde{\pi}_2^e(c)$$

about the intended kindness of the firm, where $c(w)$ is the probability measure on E prescribed by c for history w. As for the equitable payoff, let $\Pi_2(c) = \{(\pi_2(c, \check{w}), \pi_1(\check{w}, c)) | \check{w} \in \Delta W\}$ be the set of expected payoff pairs that the firm can induce if the worker follows behavioral strategy c. Then $\tilde{\pi}_2^e(c)$ is derived from the set of Pareto efficient pairs from the closure of $\Pi_2(c)$ according to assumption (A1). Observe that $\tilde{\lambda}^{IR}$ depends on the history w but no longer on the firm's behavioral strategy a_1.

Definition 1'. *A strategy profile $(\hat{a}_1, \hat{a}_2)$ is an implicit-randomization intentions-equilibrium if*
(i) $\hat{a}_1 \in \arg\max_{\check{w} \in \Delta W} \ \pi_1(\check{w}, \hat{a}_2)$, *and*
(ii) $\hat{a}_2(w) \in \arg\max_{\check{e} \in \Delta E} \ \tilde{\pi}_2(\check{e}, w) + yF(\tilde{k}(\check{e}, w), \tilde{\lambda}^{IR}(w, \hat{a}_2))$, $\forall w \in W$.

Proposition 8. *In the gift-exchange game, any implicit-randomization IE $(\hat{a}_1, \hat{a}_2)$ satisfies $\hat{a}_1 = 0$, $\hat{a}_2(0) = 0$, and $\tilde{\lambda}^{IR}(0, \hat{a}_2) \leq 0$.*

Proof. Let $(\hat{a}_1, \hat{a}_2)$ be an implicit-randomization IE. Let $W^+ \subseteq W$ be the set of wages w for which $\pi_2(\hat{a}_2, w) > \tilde{\pi}_2$ for at least one $(\tilde{\pi}_2, \tilde{\pi}_1) \in \Pi_2^E(\hat{a}_2)$, and hence $\pi_1(w, \hat{a}_2) < \tilde{\pi}_1$ by Pareto efficiency

of the pairs in $\Pi_2^E(\hat{a}_2)$. From $\hat{a}_1 \in \arg\max_{\check{w}\in\Delta W} \pi_1(\check{w}, \hat{a}_2)$ it follows that $\tilde{\pi}_1 \leq \pi_1(\hat{a}_1, \hat{a}_2)$ for all $(\tilde{\pi}_2, \tilde{\pi}_1) \in \Pi_2^E(\hat{a}_2)$, so that W^+ must have $\hat{a}_1$-measure zero. Let $W^- = W \backslash W^+$, so that W^- has $\hat{a}_1$-measure one. For any $w \in W^-$ we have, by definition, that $\pi_2(\hat{a}_2, w) = \tilde{\pi}_2(\hat{a}_2(w), w) \leq \tilde{\pi}_2$ for all $(\tilde{\pi}_2, \tilde{\pi}_1) \in \Pi_2^E(\hat{a}_2)$, and hence $\tilde{\lambda}^{IR}(w, \hat{a}_2) \leq 0$. Since $\tilde{\pi}_2(e, w)$ is strictly decreasing and $\tilde{\pi}_1(w, e)$ is strictly increasing in e, we then immediately obtain that $\hat{a}_2(w) = 0$ must hold for all $w \in W^-$. Hence we have $\pi_1(\hat{a}_1, \hat{a}_2) = -\omega$, where $\omega \geq 0$ is the expected wage under $\hat{a}_1$. It then follows that $\hat{a}_1 = 0$ must hold, since otherwise $\pi_1(\hat{a}_1, \hat{a}_2) < 0 \leq \pi_1(0, \hat{a}_2)$, a contradiction to Definition 1'(i). The remaining statements then follow from the observation that, since $\{0\}$ has $\hat{a}_1$-measure one, we have $\{0\} \subseteq W^-$. □

A.2.2 Explicit Randomization

Again, let $a_1 \in A_1 = \Delta W$ be a behavioral strategy of the firm. In contrast to the previous subsection, however, we now interpret a_1 as a conscious randomization. In the spirit of Sebald (2010), we model it as the choice of a randomization device that is observable to the worker. Hence a_1 can be thought of as the pure choice of a device which subsequently randomizes according to the probability measure a_1, and the worker observes both the chosen randomization device and its outcome. Hence the worker acts at histories $(\check{w}, w) \in \Delta W \times W$ that describe the chosen device $\check{w}$ and its outcome w, where $w \in \text{supp}(\check{w})$ must hold.[32] A behavioral strategy a_2 of the worker is therefore a function assigning a probability measure $a_2(\check{w}, w) \in \Delta E$ to every such history $(\check{w}, w)$. At a history $(\check{w}, w)$, the payoff pairs that the worker can induce depend on w but not on $\check{w}$. Specifically, they are given by $\Pi_1(w) = \{(\tilde{\pi}_1(w, \check{e}), \tilde{\pi}_2(\check{e}, w)) | \check{e} \in \Delta E\}$ as before, so kindness is

$$\tilde{k}(\check{e}, w) = \tilde{\pi}_1(w, \check{e}) - \tilde{\pi}_1^e(w)$$

exactly like in the previous subsection. Now consider the decision of the firm. If it believes the worker to play c, then it can induce the payoff pairs $\Pi_2(c) = \{(\pi_2(c, \check{w}), \pi_1(\check{w}, c)) | \check{w} \in \Delta W\}$, where the expected payoffs π_i now take into account that the worker might react differently to some wage w depending on the device $\check{w}$ that has generated it. In contrast to the previous subsection, the worker is now assumed to understand that, by making an observable choice of $\check{w}$, the firm intends to give the expected payoff $\pi_2(c, \check{w})$ to the worker, but does not specifically intend the eventual realization w. Hence we write the worker's belief about the firm's kindness in history $(\check{w}, w)$ as

$$\tilde{\lambda}^{ER}(\check{w}, c) = \pi_2(c, \check{w}) - \tilde{\pi}_2^e(c),$$

where the equitable payoff $\tilde{\pi}_2^e(c)$ is defined based on $\Pi_2(c)$ as before. Importantly, the term $\tilde{\lambda}^{ER}$ now depends on $\check{w}$ but no longer on w, in contrast to the previous subsection.

[32] Formally, the support of measure $\check{w}$, denoted $\text{supp}(\check{w})$, is defined as the set of wages $w \in W$ for which every open subset $S \subseteq W$ with $w \in S$ satisfies $\check{w}(S) > 0$.

Definition 1". *A strategy profile* $(\hat{a}_1, \hat{a}_2)$ *is an explicit-randomization intentions-equilibrium if*
(i) $\hat{a}_1 \in \arg\max_{\check{w}\in\Delta W}\ \pi_1(\check{w}, \hat{a}_2)$*, and*
(ii) $\hat{a}_2(\check{w}, w) \in \arg\max_{\check{e}\in\Delta E}\ \tilde{\pi}_2(\check{e}, w) + yF(\tilde{k}(\check{e}, w), \tilde{\lambda}^{ER}(\check{w}, \hat{a}_2))$*,* $\forall(\check{w}, w) \in \Delta W \times W$*,* $w \in supp(\check{w})$*.*

Proposition 9. *In the gift-exchange game, any explicit-randomization IE* $(\hat{a}_1, \hat{a}_2)$ *satisfies* $\hat{a}_1 = 0$*,* $\hat{a}_2(0, 0) = 0$*, and* $\tilde{\lambda}^{ER}(\hat{a}_1, \hat{a}_2) \leq 0$*.*

Proof. Let $(\hat{a}_1, \hat{a}_2)$ be an explicit-randomization IE, so $\hat{a}_1 \in \arg\max_{\check{w}\in\Delta W} \pi_1(\check{w}, \hat{a}_2)$ by definition. This implies $\pi_1(\hat{a}_1, \hat{a}_2) \geq \tilde{\pi}_1$ for all $(\tilde{\pi}_2, \tilde{\pi}_1) \in \Pi_2^E(\hat{a}_2)$. Pareto efficiency of the payoff pairs in $\Pi_2^E(\hat{a}_2)$ then also implies $\pi_2(\hat{a}_2, \hat{a}_1) \leq \tilde{\pi}_2$ for all $(\tilde{\pi}_2, \tilde{\pi}_1) \in \Pi_2^E(\hat{a}_2)$. This implies $\tilde{\lambda}^{ER}(\hat{a}_1, \hat{a}_2) \leq 0$. By definition of equilibrium, $\hat{a}_2(\hat{a}_1, w) \in \arg\max_{\check{e}\in\Delta E} \tilde{\pi}_2(\check{e}, w) + yF(\tilde{k}(\check{e}, w), \tilde{\lambda}^{ER}(\hat{a}_1, \hat{a}_2))$ must hold for all $w \in \text{supp}(\hat{a}_1)$. Since $\tilde{\lambda}^{ER}(\hat{a}_1, \hat{a}_2) \leq 0$, we again immediately obtain $\hat{a}_2(\hat{a}_1, w) = 0$ for all $w \in \text{supp}(\hat{a}_1)$. Hence we have $\pi_1(\hat{a}_1, \hat{a}_2) = -\omega$, where $\omega \geq 0$ is the expected wage under $\hat{a}_1$. It then follows that $\hat{a}_1 = 0$ must hold, since otherwise $\pi_1(\hat{a}_1, \hat{a}_2) < 0 \leq \pi_1(0, \hat{a}_2)$, a contradiction to Definition 1"(i). □

A.3 Proof of Proposition 6

Step 1. We first derive the kindness term $\tilde{k}(e, w)$. Since $\tilde{a}_2(w) = \min\{w/2, 1\}$ maximizes the worker's material payoff $\tilde{\pi}_2(e, w) = ew - e^2$, which is strictly concave in e, we obtain $\Pi_1^E(w) = \left\{\left(e(V - w), ew - e^2\right) | e \in [\min\{w/2, 1\}, 1]\right\}$. The equitable payoff when $w \in [0, 2]$ is thus $\tilde{\pi}_1^e(w) = (1/2)(V - w)((w/2) + 1)$, and $\tilde{\pi}_1^e(w) = V - w$ whenever $w \in (2, V]$. We then obtain

$$\tilde{k}(e, w) = \begin{cases} (V - w)(e - 1) & \text{if} \quad w \in (2, V], \\ (V - w)\left(e - \frac{(w/2)+1}{2}\right) & \text{if} \quad w \in [0, 2]. \end{cases}$$

Step 2. In any IE $(\hat{a}_1, \hat{a}_2)$, the worker's utility $ew - e^2 + y\tilde{k}(e, w)\tilde{\lambda}(w, \hat{a}_2)$ must be maximized by $e = \hat{a}_2(w)$ for every $w \in [0, V]$. It is easily verified that the objective is strictly concave in e (for any fixed w). The first-order condition is identical for the cases $w \in [0, 2]$ and $w \in (2, V]$ and characterizes the following effort level:

$$\bar{e}(w) = \frac{w}{2} + \frac{y\tilde{\lambda}(w, \hat{a}_2)(V - w)}{2}.$$

Concavity implies that $\hat{a}_2(w) = \bar{e}(w)$ when $\bar{e}(w) \in [0, 1]$ and $\hat{a}_2(w) = 1\,(= 0)$ when $\bar{e}(w) > 1\,(< 0)$.

Step 3. Now consider a cut-off profile $(\hat{a}_1, \hat{a}_2)$ with cut-off value $\hat{w} \in [0, V]$. We immediately obtain $\Pi_2^E(\hat{a}_2) = \{(w - 1, V - w) | \hat{w} \leq w\}$ and $\tilde{\pi}_2^e(\hat{a}_2) = ((V + \hat{w})/2) - 1$. This implies

$$\tilde{\lambda}(w, \hat{a}_2) = \begin{cases} w - \left(\frac{V+\hat{w}}{2}\right) & \text{if} \quad \hat{w} \leq w, \\ 1 - \left(\frac{V+\hat{w}}{2}\right) & \text{if} \quad w < \hat{w}. \end{cases}$$

Step 4. Optimality of $\hat{a}_2(w) = 1$ for all $w \geq \hat{w}$ now requires $\bar{e}(w) \geq 1$ for all those wages, i.e.

$$y\left[\left(\frac{V+\hat{w}}{2}\right)-w\right]\leq\frac{w-2}{V-w}$$

after substitution of $\tilde{\lambda}$ and some rearrangements. Since the LHS of this inequality is strictly decreasing and the RHS is strictly increasing in w, it is satisfied by all $w\geq\hat{w}$ if and only if it is satisfied by $w=\hat{w}$, i.e.

$$y\left[\left(\frac{V-\hat{w}}{2}\right)\right]\leq\frac{\hat{w}-2}{V-\hat{w}}. \tag{3}$$

Since the LHS of (3) is strictly decreasing and the RHS is strictly increasing in $\hat{w}$, it yields a lower bound w^l for $\hat{w}$, implicitly defined by

$$y\left[\left(\frac{V-w^l}{2}\right)\right]=\frac{w^l-2}{V-w^l}. \tag{4}$$

It also follows from this expression that $2<w^l<V$ must hold. Analogously, the condition for $\hat{a}_2(w)=0$ to be optimal for all $w<\hat{w}$ can be reduced to

$$y\left[\left(\frac{V+\hat{w}}{2}\right)-1\right]\geq\frac{\hat{w}}{V-\hat{w}}.$$

Both the LHS and the RHS of this inequality are increasing in $\hat{w}$, but it can be shown that the slope of the LHS is strictly smaller than the slope of the RHS (since condition (3) holds). Hence we implicitly obtain upper bound w^h by

$$y\left[\left(\frac{V+w^h}{2}\right)-1\right]=\frac{w^h}{V-w^h}, \tag{5}$$

which must satisfy $w^h<V$.

Step 5. It remains to be shown under which conditions $w^l\leq w^h$ holds, so that the requirements for equilibrium existence can be met simultaneously. Fix w^l as defined in (4) and suppose we evaluate (5) at the value w^l instead of w^h. Then the LHS of (5) would be (weakly) larger than the RHS if and only if $w^l\leq w^h$, by the above arguments. Dividing the LHS of (4) by $y((V+w^l)/2-1)$ and the RHS by $w^l/(V-w^l)$ thus yields

$$\frac{V-w^l}{V+w^l-2}\leq\frac{w^l-2}{w^l}$$

if and only if $w^l\leq w^h$. After some rearrangements we then obtain that $w^l\leq w^h$ if and only if $w^l\geq 1+\sqrt{V-1}$. Substituting $w^l=1+\sqrt{V-1}$ into (4) yields, after some simplifications, $y=2/((V-1)^{3/2}-(V-1))$. Inspection of (4) reveals that w^l must be strictly increasing in y, so that we have $w^l\leq w^h$ if and only if $y\geq 2/((V-1)^{3/2}-(V-1))$, which is the existence condition in the proposition. Given the cut-off strategy $\hat{a}_2$, the fact that $w=\hat{w}$ (uniquely) maximizes $\tilde{\pi}_1(w,\hat{a}_2(w))$ as required in the definition of IE is immediate.